G-Cloud 15 Amazon Web Services EMEA SARL, UK Branch (AWS)

AWS Services

Service Definition Document

Lot 1a Infrastructure as a Service (IaaS) and Platform as a Service (PaaS)

Lot 2a Infrastructure as a Service (I-SaaS)

Lot 2b Software as a Service (SaaS)

30 January 2026

Individual AWS service definitions included in Annex

G-Cloud 15 Amazon Web Services EMEA SARL, UK Branch (AWS)

AWS Services

Service Definition Document

Lot 1a Infrastructure as a Service (IaaS) and Platform as a Service (PaaS)

Lot 2a Infrastructure as a Service (I-SaaS)

Lot 2b Software as a Service (SaaS)

Please insert the appropriate legal disclaimer using the WWPS Disclaimers and Confidentiality Legend located in Pathfinder.

Table of Contents

INTRODUCTION 1
CROSS-SERVICE DEFINITIONS 1
1 Shared Responsibility Model 1
2 Technical Architecture & Service Delivery 5
2.1 Infrastructure Details 5
2.2 Service Delivery & Access 7
2.3 Service Granularity & Measurement 8
3 Pricing & Commercial Model 9
4 Operational Capabilities 11
4.1 Service Management 11
4.2 Data Protection 13
5 Support & User Enablement 14
6 Service Continuity & Exit 16
7 Innovation & Technology Advancement 20
8 Standards Compliance & Sustainability 21
9 Annex 24

INTRODUCTION

Please note that we have consolidated common elements of each Service Offering and have provided descriptions for these common elements that apply equally to each Service Offering—see "CROSS-SERVICE DEFINITIONS". To find out more about AWS on G-Cloud and AWS Cloud services, visit us at AWS on G-Cloud UK.

The AWS Free Tier enables you to gain free, hands-on experience with AWS products and services. The AWS Free Tier provides customers the ability to explore and try out AWS services free of charge up to specified limits for each service, with offerings available for up to 12 months for traditional accounts or through a $200 credit system (valid for 6 months) for accounts created after July 15, 2025.. See http://aws.amazon.com/free/ for service-specific details.

CROSS-SERVICE DEFINITIONS

1 Shared Responsibility Model

Customers and AWS share security and compliance responsibilities. As shown in Figure 1, AWS operates the services and underlying infrastructure that make up the cloud. We secure both the facilities that house the hardware and the foundational cloud technologies, like virtualization tools. You are responsible for managing your data, guest operating systems, and applications, as well as configuring the security of your AWS environment.

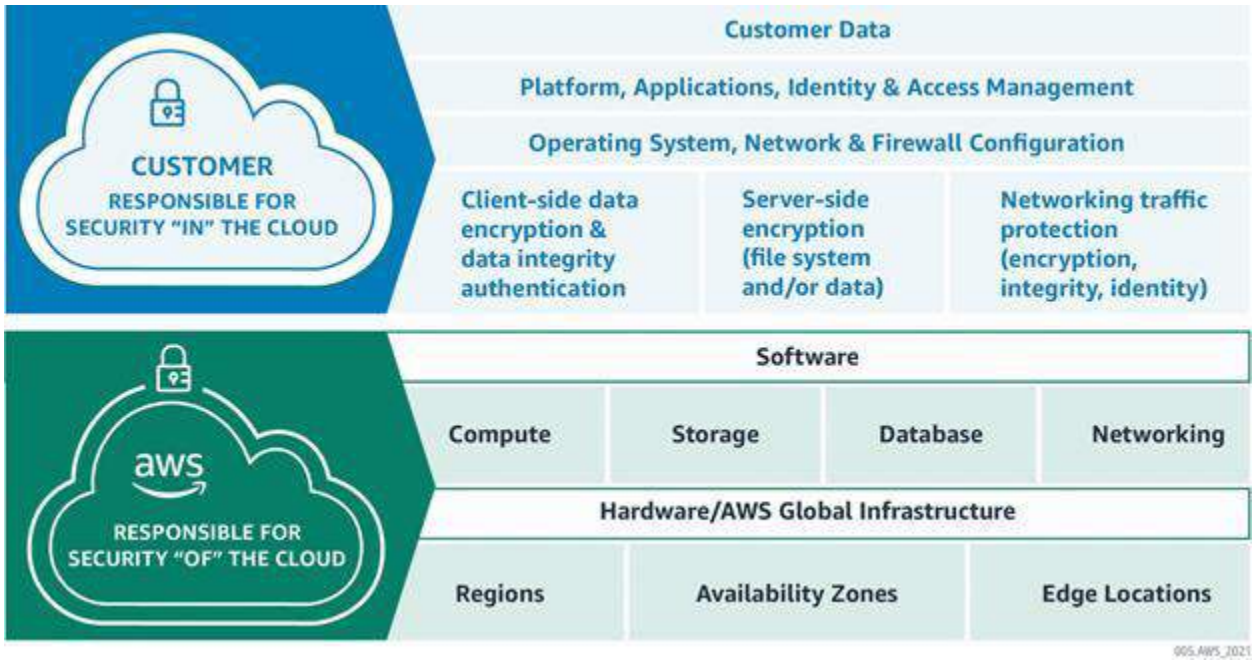

Figure 1. The Shared Responsibility Model.

This model establishes that AWS is responsible for the underlying cloud infrastructure, while customers are responsible for their data and applications. The customer's specific responsibilities depend on the AWS services deployed.

AWS Security Responsibilities: Security of the Cloud

AWS is responsible for *security of the cloud*. This means protecting the infrastructure that runs all of the services we offer. This infrastructure is composed of hardware, software, networking, and facilities. To protect that infrastructure, we designed and built it with the unique needs of the cloud in mind.

We use redundant and layered controls, continuous validation and testing, and automation to monitor and protect our underlying infrastructure 24/7. We replicate these controls in every new data center or service. We built our data centers and network architecture to meet the requirements of our most security-sensitive customers. This includes the Intelligence Community, DoD, and civilian agencies, as well as organizations in regulated industries like healthcare, aerospace and defense, energy, and financial services. As a result, all our customers get a resilient infrastructure designed for high security.

AWS has implemented sophisticated technical and physical measures against unauthorized access to protect our customers' security. Customers can validate the security controls in place within the AWS environment through our certifications and reports, including the AWS System and Organization Control (SOC) 1, 2, and 3 reports; International Organization for Standardization (ISO) 27001, 27017, 27018, and 9001 certifications; and Payment Card Industry (PCI) Data Security Standard (DSS) compliance reports. The ISO 27018 certification demonstrates that we have a system of controls in place for the privacy protection of customer content. Independent third-party auditors produce these certifications and reports, which attest to the design and operating effectiveness of AWS security controls. You can read more about AWS compliance certifications, reports, and alignment with industry practices and standards at http://aws.amazon.com/compliance/.

Customer Responsibilities: Security in the Cloud

AWS customers are responsible for *security in the cloud*. This refers to your applications, content, and systems. You have complete control over which services you use and whom you empower to access your content and services, including what credentials are required to gain access. When using AWS services, you retain ownership and control of your content.

Because you, rather than AWS, control these important factors, you are responsible for securing any content that you store or process using our services and any content you connect to AWS infrastructure. This means you are responsible for the guest operating system, applications on your compute instances, and content stored and processed in AWS storage, databases, or other services. You will need to set up access control policies so that only authorized individuals can access resources.

Security Responsibilities by Service Layer

The division of responsibilities varies significantly based on the AWS service layer selected.

For Infrastructure as a Service (IaaS) offering like Amazon EC2, customers have responsibilities including guest operating system management, application software and utilities, security group configuration, network access control lists, identity and access management, and data encryption.

For Platform as a Service (PaaS) offerings such as managed services, AWS assumes additional responsibilities including operating system maintenance and patching, database or platform software updates, network configuration and security, backup and recovery infrastructure, and platform-level security controls. Customers focus on data management, encryption options, IAM policies, and application-level access controls.

For serverless services like AWS Lambda, AWS manages the most comprehensive set of responsibilities including infrastructure and foundation services, operating system and runtime environment, application platform and execution environment, server provisioning and capacity management, and network configuration and security. This frees customers to focus on securing application code, storing and accessing sensitive data, implementing identity and access management, and maintaining monitoring and logging.

Shared Responsibility and Operational Controls

Just as AWS and our customers share responsibilities to operate the IT environment, we also share responsibilities to manage, operate, and verify IT controls. As every customer's deployment in AWS is different, customers can shift management of certain controls to AWS, resulting in a distributed control environment. The following are examples of controls managed by AWS, our customers, or both AWS and customers.

Inherited Controls

We reduce your security burden by managing the controls associated with AWS's physical infrastructure (inherited controls).

Shared Controls

Some controls apply to both the infrastructure layer (AWS responsibility) and customer layers (customer responsibility), but in completely separate contexts or perspectives (shared controls). With shared controls, AWS provides the requirements for the infrastructure, and the customer provides their own control implementation within their AWS services. Examples of these shared controls include the following:

- **Patch Management:** AWS is responsible for patching and fixing flaws within the infrastructure, but customers are responsible for patching their guest operating system and applications.
- **Configuration Management:** AWS maintains the configuration of our infrastructure devices, but customers are responsible for configuring their own guest operating systems, databases, and applications.
- **Awareness and Training:** AWS trains AWS employees, but customers must train their own employees.

Customer-Specific Controls

Some controls are the sole responsibility of the customer based on the applications they deploy within the AWS Cloud (customer-specific controls). Examples include service and communications protection or zone security, which may require a customer to route or zone data within specific security environments. Customers also control how they use their account and what content moves into and out of the account. They also need visibility into vulnerabilities that can threaten their applications, content, and systems.

Operational Responsibilities Definition

Beyond security, the shared responsibility model encompasses operational responsibilities. AWS manages the global infrastructure, service availability, and platform maintenance, while customers manage their workload configuration, data governance, access management, and application performance optimization. This specialization allows AWS and customers to focus on their respective areas of expertise, creating measurable outcomes across security, operations, and cost management while maintaining customer control over critical business decisions.

Shared Responsibility and Customer Data

AWS classifies customer data into two categories: customer content and account information.

Customer Content

With AWS, customers get the ability to store their content on cloud infrastructure and access it from almost anywhere over the internet. As a customer, you maintain full control of your content that you upload to the AWS services under your AWS account. This means that you

maintain responsibility for configuring access to AWS services and resources. We provide an advanced set of access, encryption, and logging features to help you do this effectively. This includes APIs through which you can configure access control permissions for any of the services you develop or deploy in an AWS environment. We do not access or use your content for any purpose without your agreement. We never use your content or derive information from it for marketing or advertising purposes.

We define ***customer content*** as software (including machine images), data, text, audio, video, or images that a customer or any end user transfers to us for processing, storage, or hosting by AWS services. For example, customer content includes content that a customer or any end user stores in Amazon Simple Storage Service (S3). Customer content does not include account information which we describe below. Customer content also does not include information included in resource identifiers, metadata tags, usage policies, permissions, and similar items related to the management of AWS resources. The terms of the AWS Customer Agreement and the AWS Service Terms apply to customer content.

As an AWS customer—and in keeping with the shared responsibility model for cloud—you maintain ownership and control over your content. AWS provides tools and guidance to support you in owning and managing your data. With these tools, you can:

- Determine where your content will be stored, including the type of storage and geographic region of that storage.
- Choose the secured state of your content—we offer customers strong encryption for your content in transit and at rest, and we provide you with the option to manage your own encryption keys.
- Manage access to your content and access to AWS services and resources through users, groups, permissions, and credentials that you control.

Legal Requests for Customer Content

AWS is vigilant about customer privacy. We will not disclose customer content unless we are required to do so to comply with the law or binding order of a governmental body. If a governmental body sends AWS a demand for customer content, we will attempt to redirect the governmental body to request that data directly from the customer. If compelled to disclose customer content to a governmental body, we will give customers reasonable notice of the demand to allow the customer to seek a protective order or other appropriate remedy unless AWS is legally prohibited from doing so. Governmental and regulatory bodies need to follow the applicable legal process to obtain valid and binding orders. We review all orders and object to overbroad or otherwise inappropriate ones. Unless prohibited from doing so, AWS notifies customers before disclosing customer content so they can seek protection from disclosure. AWS customers can encrypt their customer content, and we provide customers with the option to manage their own encryption keys.

Amazon, our parent company, knows that customers care deeply about privacy and data security and regularly publishes Amazon Information Request Reports about the types and volume of information requests that it receives, including the following:

- Amazon does not disclose customer information in response to government demands unless it is required to do so to comply with a legally valid and binding order. Unless prohibited from doing so or there is clear indication of illegal conduct in connection with the use of Amazon products or services, Amazon notifies customers before disclosing content information.
- Where Amazon needs to act to protect customers, it does. Amazon has repeatedly challenged government demands for customer information that were considered overbroad, winning decisions that have helped to set the legal standards for protecting customer speech and privacy interests. Amazon also advocates in Congress to modernize outdated privacy laws to require law enforcement to obtain a

search warrant from a court to get the content of customer communications. That is the appropriate standard, and it is the standard Amazon and its subsidiaries follow.

- While Amazon recognizes the legitimate needs of law enforcement agencies to investigate criminal and terrorist activity and cooperate with them when they observe legal safeguards for conducting such investigations, it opposes legislation mandating or prohibiting security or encryption technologies that would weaken the security of products, systems, or services its customers use, whether they be individual consumers or business customers. For AWS customers, we offer strong encryption as one of many standard security features, and we provide them the option to manage their own encryption keys. We publish security best practices documents on our website and encourage our clients to use these measures to protect sensitive content.
- Amazon is a member of numerous associations focused on protecting privacy and security, and AWS in particular has achieved a number of internationally recognized certifications and accreditations demonstrating compliance with third-party assurance frameworks. AWS clients have control over their content and where it resides.

Amazon Information Request Reports dating back to 2015 can be found on Amazon's Law Enforcement Information Requests page.

Account Information

We define account information as information about a customer that a customer provides to us in connection with the creation or administration of a customer account. For example, account information includes names, usernames, phone numbers, email addresses, and billing information associated with a customer account. The information practices described in the AWS Privacy Notice apply to account information.

Customer Virtual Instances

Customer virtual instances are solely controlled by the customer. AWS personnel do not have the ability to log into customer instances. AWS customers manage the creation and deletion of their data on AWS, maintain control of access permissions, and manage appropriate data retention policies and procedures.

Documentation To Support Your Security Objectives

AWS can help you handle your share of security of the infrastructure. We equip customers with documentation for all our services to guide you through proper configuration for security. We also offer hundreds of tools and features to help you meet your security objectives, including services for network security, configuration management, access control, and data encryption. Additionally, we offer integrated Partner solutions and support in the form of AWS customer enablement services to help you design, implement, and operate your security environment. For further information, refer to Shared Responsibilities at https://docs.aws.amazon.com/wellarchitected/latest/security-pillar/shared-responsibility.html

2 Technical Architecture & Service Delivery

2.1 Infrastructure Details

Cloud deployment model

Based on the nature of your application and the underlying services (compute, storage, database, etc.) it requires, you can use AWS services to create a flexible deployment model—a cloud based application, a hybrid deployment, or a private cloud (on premises)—that you can tailor to fit the needs of both your application and your organization. AWS

supports all four NIST-defined cloud deployment models, providing organizations with flexible infrastructure options aligned with their specific requirements and compliance needs.

Public Cloud. AWS cloud infrastructure is provisioned for open use by the public and exists on AWS premises. AWS cloud delivers massive scale, continuous innovation, and cost efficiency that traditional private cloud implementations cannot match. Through services like Amazon VPC, organizations create logically isolated environments with complete control over their virtual networking, achieving the privacy benefits of dedicated infrastructure while accessing the full breadth of AWS capabilities.

Private Cloud. AWS enables private cloud deployments through AWS Outposts, which extends AWS infrastructure, services, APIs, and tools to virtually any data center, colocation space, or on-premises facility for a truly consistent hybrid experience. Private cloud infrastructure is provisioned for exclusive use by a single organization providing enhanced control over security, compliance, and configuration while maintaining consistent AWS APIs and tools.

Hybrid Cloud. AWS provides hybrid cloud solutions that compose two or more distinct cloud infrastructures bound by standardized technology enabling data and application portability. Key hybrid services include AWS Direct Connect for dedicated network connections, AWS Outposts for on-premises AWS infrastructure, and Amazon ECS Anywhere for container orchestration across environments. These solutions bridge existing investments with cloud capabilities, supporting data center extension, application migration, and local data requirements.

Community Cloud. AWS supports community cloud models through AWS Organizations for centralized management across multiple accounts with shared governance requirements. The AWS European Sovereign Cloud represents a significant advancement in community cloud capabilities, designed specifically for European government organizations and highly regulated industries. This independent cloud infrastructure will be physically and logically separated from other AWS Regions, operating within the EU with its own separate authentication, operations, and control systems. The European Sovereign Cloud will offer the same AWS services, features, and security standards while meeting the highest levels of sovereignty requirements for regulated industries and public sector organizations in Europe.

Infrastructure location

AWS offers 240+ fully featured services from data centers globally, refer to https://aws.amazon.com/about-aws/global-infrastructure/. The AWS Cloud spans 120 Availability Zones within 38 Geographic Regions, including the Europe (London) Region launched in 2016 with 3 Availability Zones. Whether you need to deploy your application workloads across the globe in a single click, or you want to build and deploy specific applications closer to your end users with single-digit millisecond latency, AWS provides cloud infrastructure where and when you need it.

AWS Implementation of NIST 5 Essential Characteristics

AWS implements all NIST 5 Essential Characteristics through specific technical capabilities that demonstrate comprehensive cloud computing functionality.

On-Demand Self-Service. AWS enables consumers to unilaterally provision computing capabilities automatically without requiring human interaction. AWS allocates resources using API calls, reducing provisioning time to minutes. Access methods include the AWS Management Console for web-based interface capabilities, AWS APIs for direct programmatic access, and the AWS Command Line Interface for unified command-line management.

Broad Network Access. AWS implements broad network access through global infrastructure comprising over 400 Points of Presence across more than 300 cities worldwide. The global infrastructure provides low latency and high network quality through a fully redundant 100 GbE fiber network backbone. Services like Amazon CloudFront, AWS Wavelength, and AWS Local Zones extend network access capabilities to edge locations and 5G networks.

Resource Pooling. AWS resource pooling uses multi-tenancy where physical and virtual resources are dynamically assigned according to consumer demand. AWS has massive economies of scale, enabling virtually unlimited resources for less cost than maintaining independent data centers. Multi-tenant architecture patterns include Silo, Pool, and Bridge models, with services like Amazon DynamoDB and AWS Lambda optimized for elastic resource sharing.

Rapid Elasticity. AWS rapid elasticity operates through Amazon EC2 Auto Scaling, which automatically launches or terminates instances according to predefined conditions. Organizations can deploy hundreds to thousands of servers in minutes to meet workload demands. Elasticity helps drastically reduce spending by matching resource utilization to demand at the lowest possible cost. AWS Lambda provides built-in elastic scalability, scaling automatically from few requests to thousands per second.

Measured Service. AWS implements a measured service through pay-as-you-go pricing where customers pay only for computing resources consumed. Amazon CloudWatch serves as the core monitoring component, collecting data from over 70 AWS services. AWS generates detailed billing reports and provides cost management tools including AWS Cost Explorer and AWS Budgets for tracking usage and costs.

The NIST Cyber Security Framework (CSF). The whitepaper, 'Aligning to the NIST CSF in the AWS Cloud', helps you understand how AWS cloud offerings align to the NIST CSF, and the role of third-party validations confirming AWS services' alignment with the NIST CSF risk management practices. You can access the whitepaper at https://d0.awsstatic.com/whitepapers/compliance/NIST_Cybersecurity_Framework_CSF_Jan_2025.pdf

2.2 Service Delivery & Access

Service Delivery

AWS delivers its cloud services primarily through secure connections over the public internet, using HTTPS/TLS encryption to protect data in transit and approved networks.

The AWS Global Network offers multiple layers of encryption for data in transit, digitally signed routing information, and unique threat intelligence. UK government organizations can connect through networks including PSN, HSCN, and Janet.

For UK government approved networks, AWS supports connectivity options through AWS Direct Connect, which establishes dedicated network connections from premises to AWS.

AWS VPN solutions meet NCSC Foundation profile requirements and can be used with systems running at OFFICIAL classification level. AWS PrivateLink provides private connectivity between VPCs, AWS services, and on-premises applications.

Supplier Portal and Interface Accessibility

AWS demonstrates strong commitment to accessibility compliance through systematic evaluation against WCAG 2.1 and 2.2 Level AA standards, Section 508, and EN 301 549. AWS has received Accessibility Conformance Reports (ACRs) for over 160 AWS services as

of 2024. These ACRs provide assessments against WCAG Level A and AA requirements and are available through AWS Artifact.

AWS demonstrates a broader commitment to accessibility across its products—focusing on keyboard operability, color contrast, accessible labels, and screen reader compatibility. AWS generally integrates accessibility from the design phase and works with experts to help ensure accessible code. AWS customers can use almost all services directly without retrofitting or adjusting them for accessibility. The AWS Command Line Interface (CLI) can be used by customers who use assistive technology such as Job Access with Speech (JAWS), NonVisual Desktop Access (NVDA), and alternative input devices. Throughout the service development cycle, AWS developers use automated accessibility testing tools and manual testing with assistive technologies.

Multi-cloud Support

At AWS, we give customers the freedom to choose technology that best suits their needs.

For example, you can use the same service (AWS Systems Manager) to patch and update Amazon Elastic Compute Cloud (Amazon EC2) instances, servers running on-premises, and servers provided by other cloud providers. Similarly, you can use Amazon CloudWatch to monitor applications, compute resources, and other cloud resources in all those environments.

Additionally, Oracle Database@AWS allows organizations to run Oracle workloads on AWS while maintaining compatibility with Oracle Cloud Infrastructure, creating a bridge between cloud platforms. For data movement across clouds, AWS DataSync supports transfers between AWS and other providers including Google Cloud, Azure, and Oracle Cloud, allowing consistent data access regardless of location.

Security remains paramount in multi-cloud environments, with Amazon Security Lake centralizing security data from AWS, Azure, GCP, and on-premises sources into a unified data lake using the Open Cybersecurity Schema Framework. For comprehensive visibility, Amazon CloudWatch extends monitoring capabilities across clouds, allowing organizations to query metrics from AWS, Azure Monitor, and other environments through a single interface.

AWS's approach focuses on practical solutions that address real multi-cloud challenges while applying AWS's strengths in security, scalability, and operational excellence. The AWS Solutions for Hybrid and Multicloud page contains additional examples of our extension-based approach to adding new capabilities as well as documentation on effective multi-cloud strategies.

We allow free data transfer out to the internet (DTO) when you want to move outside of AWS. We are committed to helping you be successful regardless of your approach.

We work closely with customers to develop a deep understanding of your multi-cloud goals. We can partner with you to help build the business case for your multi-cloud strategy and guidance on best practices. We also offer hands-on multicloud training for builders. We provide prescriptive guidance and professional services to set up and operate a Cloud Center of Excellence (CCOE) for a multi-cloud environment.

2.3 Service Granularity & Measurement

Metrics and Reporting

AWS provides flexible metric reporting through Amazon CloudWatch with granularity options tailored to different monitoring needs. The service offers two primary monitoring tiers: Basic Monitoring, which collects metrics at 5-minute intervals at no additional charge, and Detailed

Monitoring, which provides enhanced 1-minute interval metrics for an additional fee. Service-specific granularity varies across the AWS environments—Amazon S3 offers daily storage metrics for free while providing optional 1-minute request metrics at additional cost. For details, refer to ttps://docs.aws.amazon.com/AmazonS3/latest/userguide/cloudwatch-monitoring.html

CloudWatch's tiered data retention policy keeps 1-minute data points for 15 days, 5-minute data for 63 days, and hourly metrics for 455 days (15 months), allowing both real-time operational monitoring and long-term trend analysis. For details, refer to https://docs.aws.amazon.com/ec2/latest/devguide/monitor.html This multi-tiered approach allows organizations to balance monitoring precision with cost considerations across their AWS infrastructure.

Additionally, monitoring services include Amazon EventBridge for viewing AWS Backup events, AWS CloudTrail for monitoring API calls, and Amazon SNS for event notifications

Billing and Usage

AWS offers comprehensive billing transparency through multiple exportable formats and proactive notification systems. The AWS Cost and Usage Report delivers detailed billing data in CSV format, downloadable directly from the Billing Console or automatically delivered to an Amazon S3 bucket for integration with analytics tools. AWS Cost Explorer enables viewing and analyzing costs and usage with up to 12 months of historical data and 12-month forecasting capabilities.

For usage monitoring, AWS provides multi-layered notification capabilities through CloudWatch Billing Alarms that trigger when spending approaches or exceeds predefined thresholds. These alerts can be delivered via email, SMS, or through other AWS services. AWS Budgets further enhances cost control by setting custom budgets with email or SNS notifications when thresholds are exceeded, while AWS Cost Explorer generates downloadable reports with customizable filters for detailed spending analysis. Together, these tools provide financial visibility and proactive cost management across all AWS services.

Service Level Agreements

AWS maintains Service Level Agreements (SLAs) for all its services, documenting specific up-time commitments and service credit policies. The central repository for all AWS SLAs is accessible at https://aws.amazon.com/legal/service-level-agreements/, where customers can review detailed terms for each service.

3 Pricing & Commercial Model

AWS offers comprehensive pricing structures for diverse organizational needs while supporting UK government procurement requirements through the G-Cloud 15 framework. The pricing model encompasses multiple options from pay-as-you-go flexibility to long-term commitment savings, all available in GBP currency with transparent cost optimization tools.

Pricing Structure Overview

AWS's foundational pricing model operates on a pay-as-you-go basis, allowing customers to pay only for resources consumed without long-term commitments. On-Demand Instances enable payment for compute capacity by the hour or second with no long-term commitments, transforming large, fixed costs into smaller variable costs. This approach allows organizations to adapt to changing business needs without overcommitting budgets and improves responsiveness to changes.

Most workloads start as on-demand, but alternative pricing constructs can provide additional discounts based on utilization and capacity needs. AWS provides multiple pricing models to optimize costs based on workload characteristics and organizational requirements.

Savings Plans Framework. AWS Savings Plans serve as a primary cost optimization tool, offering flexible pricing models that provide significant savings compared to On-Demand pricing. AWS offers multiple Savings Plans types: Compute Savings Plans provide up to 66% savings and automatically apply to EC2, Lambda, and Fargate usage regardless of instance family, size, region, or operating system. EC2 Instance Savings Plans offer the lowest prices with savings up to 72% for commitment to specific instance families in a region. Database Savings Plans provide up to 35% savings on AWS database services with 1-year commitments. SageMaker Savings Plans offer up to 64% savings on Amazon SageMaker usage.

Volume Discounts and Enterprise Benefits

AWS provides volume discounts as usage grows, with customers able to take advantage of up to 72% discounts when committing to Savings Plans or Reserved Instances for predictable usage. Purchasing large numbers of Amazon EC2 Reserved Instances in an AWS Region provides discounts on upfront fees and hourly fees for future RI purchases. Enterprise customers benefit from consolidated billing through AWS Organizations, allowing multiple accounts to benefit from volume discounts and shared Reserved Instance benefits.

Pay-as-You-Go Availability

AWS offers you a pay-as-you-go approach for pricing for the majority of our cloud services. With AWS you pay only for the individual services you need, for as long as you use them, and without requiring long-term contracts or complex licensing (though Reserved Instances offer optional minimum commitments for cost savings). The pay-as-you-go model helps organizations to scale resources dynamically without upfront investments, supporting the agility requirements of modern government operations.

GBP Currency Support

AWS supports billing in British Pounds (GBP) and other local currencies, allowing UK-based organizations to manage AWS costs in their preferred currency for better financial planning and reporting.

Cost Optimization Tools and Mechanisms

AWS provides comprehensive cost optimization tools including AWS Cost Explorer, which allows viewing and analyzing costs and usage with up to 12 months of historical data and 12-month forecasting. AWS Cost Optimization Hub quantifies estimated cost savings of recommendations, incorporating specific AWS pricing and discounts such as Reserved Instances and Savings Plans. AWS Compute Optimizer uses machine learning to analyze historical utilization metrics and recommend optimal resource configurations.

AWS follows established cost optimization principles that apply across nearly all environments, focusing on rightsizing resources to match actual needs, increasing elasticity by turning off resources when not needed, using optimal pricing models based on workload characteristics, optimizing storage by selecting appropriate tiers, and measuring, monitoring, and improving through cost allocation tagging and continuous monitoring.

FOCUS 1.2 Standards Commitment

AWS achieved general availability of FOCUS 1.2 on November 19, 2025, allowing standardized cost and usage data across multi-cloud environments with enhanced capabilities for invoice reconciliation, capacity reservation tracking, and improved data

normalization.[1] AWS serves as a steering committee member and contributor to the FOCUS project, having been actively involved in FOCUS specification development since joining the FinOps Foundation as a premier member in October 2023.

FOCUS 1.2 includes significant enhancements over previous versions, with 14 additional columns compared to FOCUS 1.0, supporting hourly, daily, and monthly time granularity, and new features including InvoiceId for direct correlation with AWS invoices, CapacityReservationId and CapacityReservationStatus columns for tracking reserved capacity, and enhanced integration capabilities with SaaS provider FOCUS 1.2 datasets. Organizations can access AWS Data Exports for FOCUS 1.2 through the AWS Billing and Cost Management console.

Spot Instances for Cost-Sensitive Workloads

For fault-tolerant workloads, AWS Spot Instances provide access to unused Amazon EC2 capacity at up to 90% discount compared to on-demand prices. This option enables significant cost reductions for appropriate workload types while maintaining performance requirements.

Public Visibility and Transparency

AWS maintains publicly visible pricing at https://aws.amazon.com/pricing for all services, providing transparency for accurate cost planning. The AWS Pricing Calculator allows estimation of costs for running AWS services, supporting informed procurement decisions. This transparency aligns with G-Cloud 15 requirements for clear pricing visibility and supports competitive procurement processes.

4 Operational Capabilities

4.1 Service Management

AWS provides service management capabilities through various integrated tools and services that help with backups, scaling, monitoring, and cost management.

Backups and Recovery

AWS Backup serves as the centralized, fully managed service that automates data protection across AWS services and hybrid workloads. Using the AWS global infrastructure, AWS Backup is designed to achieve durability of 99.999999999% (11 nines). The service provides centralized console management, automated backup scheduling, backup retention management, and comprehensive monitoring and alerting capabilities.

AWS Backup integrates with Amazon CloudWatch to provide monitoring capabilities, with dashboard displays showing job metrics by count and customizable time frames ranging from 1 hour to 1 week. The monitoring system includes Amazon EventBridge for viewing AWS Backup events, AWS CloudTrail for monitoring API calls with source IP and user identification, and Amazon SNS for subscribing to backup, restore, and copy events.

AWS Backup Audit Manager provides continuous evaluation of backup activity and generates audit reports that help demonstrate compliance with regulatory requirements. Third-party auditors assess the security and compliance of AWS Backup as part of multiple AWS compliance programs, including SOC, PCI, FedRAMP, HIPAA, and others.

[1] FOCUS (FinOps Open Cost and Usage Specification) is an open specification supported by the FinOps Foundation that standardizes cost and usage data.

Amazon Data Lifecycle Manager automates the creation, retention, and deletion of Amazon EBS snapshots, helping you protect valuable data through regular backup schedules, maintain compliance with retention requirements, and reduce storage costs by removing outdated backups. This service, combined with Amazon CloudWatch and Amazon CloudTrail, provides a complete backup solution for EBS volumes at no additional cost.

For Windows environments, you can create VSS application-consistent snapshots using AWS CLI, PowerShell, or the AWSEC2-ManageVssIO SSM Document, which helps ensure database integrity during backups. AWS also supports manual snapshot management, though automated solutions are recommended. For details, refer to https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/application-consistent-snapshots.html

Scaling Capabilities

AWS provides multiple scaling options. Amazon EC2 Auto Scaling maintains application availability by automatically adding or removing EC2 instances according to the conditions you define. It can detect when instances or applications are unhealthy and replace them automatically to maintain availability.

For Spot Fleet instances, AWS offers automatic scaling through integration with Amazon CloudWatch and Application Auto Scaling. You can implement step scaling policies that respond to CloudWatch alarms or target tracking policies that maintain specific metric values. AWS Lambda also supports automatic scaling with provisioned concurrency. Using Application Auto Scaling, you can create target tracking scaling policies that adjust provisioned concurrency levels based on utilization metrics.

Usage Monitoring

Amazon CloudWatch is the primary service for monitoring AWS resources and applications. It collects and tracks metrics, monitors log files, sets alarms, and automatically reacts to changes in your AWS environment.

The CloudWatch dashboard shows current alarms and status, graphs of alarms and resources, and service health status. It allows you to graph monitoring data to troubleshoot issues, search and browse metrics, create and edit alarms, and view at-a-glance overviews of your resources. For EC2 instances, CloudWatch metrics can be aggregated by Auto Scaling group, allowing you to view statistics across all instances in a group.

Overspend Alerts

AWS Budgets allows organizations to set custom budgets to track costs and usage, with alerts triggered when actual or forecasted costs exceed defined thresholds. AWS Cost Anomaly Detection uses machine learning to continuously monitor cost and usage patterns, automatically learning normal spending behavior and alerting only when genuine anomalies occur. Amazon CloudWatch creates billing alerts that notify users when AWS service usage exceeds defined thresholds, with email notifications sent when usage exceeds specified amounts. The system analyzes spending patterns within an organisational context, calibrating anomaly detection according to how AWS accounts are organized by department, project, or functional area.

Organisations can implement cost control systems focusing on four main components: granting appropriate permissions to users, budgeting spend with custom thresholds, monitoring and analyzing cost progression toward limits, and taking actions to reduce unintentional costs.

4.2 Data Protection

Data Protection Between Buyer and Supplier

AWS implements comprehensive data protection measures across its services to secure data in transit, at rest, and during processing. These measures are part of AWS's shared responsibility model, where AWS secures the infrastructure while customers maintain responsibility for their data security configurations.

AWS data encryption capabilities span three critical protection states: encryption at rest, encryption in transit, and encryption in use. AWS Key Management Service integrates with over 120 AWS services, providing encryption capabilities across compute, storage, database, and networking services. All traffic between AWS data centers is transparently encrypted at the physical layer before it leaves AWS-secured facilities. All traffic within the AWS network is encrypted by default, and network traffic within virtual private clouds cannot be spoofed or sniffed. AWS CloudHSM provides dedicated HSM instances validated at FIPS 140-2 Level 3, critical for demonstrating compliance with security and privacy regulations including PCI DSS, GDPR, HIPAA, and FedRAMP.

Data Protection Within Supplier Network

AWS implements defense-in-depth encryption strategies across multiple layers including physical and perimeter security with transparent encryption between data centers, internal network encryption by default, host-level security through the AWS Nitro System, and data protection. The AWS Nitro System provides enhanced security through hardware-based isolation and dedicated security chips.

AWS supports 143 security standards and compliance certifications, including PCI-DSS, HIPAA/HITECH, FedRAMP, GDPR, FIPS 140-3, and NIST 800-171, helping customers satisfy compliance requirements around the globe. For UK public sector organizations, AWS is compliant with NHS Digital Data Security and Protection Toolkit, ISO 27001/27017/27018, Cyber Essentials Plus, GDPR/Data Protection Act 2018, and NCSC Cloud Security Principles.

Connected Public Sector Networks

AWS supports UK government networks through dedicated connectivity options. AWS Direct Connect delivers data through a private network connection between facilities and AWS, bypassing the public internet for secure access to workloads. The AWS VPN configuration meets NCSC Foundation profile requirements and can be used with systems running at OFFICIAL classification level. For NHS organizations specifically, AWS is connected to HSCN through several partner organizations, allowing NHS Trusts to access AWS Cloud services through HSCN if desired. Additionally, services such as AWS PrivateLink provide private connectivity between VPCs and AWS services without exposing traffic to the public internet, reducing exposure to brute force and DDoS attacks.

The AWS EU (London) Region has achieved Public Services Network (PSN) assurance, allowing UK public sector organizations to connect through PSN-certified AWS Direct Connect partners. This certification demonstrates that AWS meets PSN's stringent requirements for handling UK OFFICIAL classified data. AWS infrastructure in the London Region is PSN-assured, so customers don't need to include AWS infrastructure in their own PSN certification process.

5 Support & User Enablement

User Support

The AWS Support Center serves as the cornerstone of our support infrastructure, where we provide a unified portal for users to access documentation, submit support tickets, and monitor the health of AWS services in real-time. AWS offers a tiered support model that scales with business needs. This includes Business Support+, Enterprise Support, and Unified Operations premier support plans. The three primary support plans are designed to accelerate cloud operations. This is complemented by AWS community forums at https://repost.aws/ where developers and architects freely exchange knowledge, troubleshooting tips, and implementation strategies.

Business Support+ serves as the minimum recommended plan for production workloads, providing response times of less than 30 minutes for business/mission-critical system issues, less than 1 hour for production system issues, and less than 4 hours for production system impairment. Pricing follows a tiered structure with a minimum of £29 per month per account or percentage-based fees: 9% of monthly AWS charges up to £10K, 7% from £10K-£80K, 5% from £80K-£250K, and 3% over £250K.

Enterprise Support is recommended for business-critical workloads across organizations seeking expert guidance offering response times as fast as 15 minutes for business/mission-critical issues. This plan includes designated Technical Account Managers who combine deep AWS knowledge with understanding of business objectives, plus proactive services including AWS Well-Architected Reviews, architectural guidance, and planned events support. Pricing requires a minimum of £5K per month or tiered percentages: 10% up to £150K, 7% from £150K-£500K, 5% from £500K-£1M, and 3% over £1M.

Unified Operations targets mission-critical workloads requiring enhanced resilience and application-specific expertise, providing response times as fast as 5 minutes from Incident Management Engineers for business-critical issues. This premium tier includes 24/7 workload monitoring, designated Domain Specialist Engineers, and Incident Management Engineers, with AWS Countdown Premium and AWS Incident Detection and Response included at no additional cost. Pricing requires a minimum of £50K per month or tiered percentages: 10% up to £1M, 6% from £1M-£5M, and 5% over £5M.

All Premium Support plans include unlimited 24/7 contextual recommendations through AI-powered troubleshooting, providing instant, context-aware assistance that understands specific customer environments. Support channels encompass 24/7 phone, web, chat, and email access, AWS Support App in Slack integration, unlimited cases and contacts, third-party software support, and AWS Support API access. All plans are billed monthly with no long-term contracts, though customers must wait a minimum of 30 days to cancel subscriptions.

User Support Accessibility

AWS provides multiple accessible communication channels to accommodate different user needs and preferences. The AWS Command Line Interface serves as a critical accessibility accommodation, enabling users with disabilities to access services through assistive technology such as Job Access with Speech (JAWS), Nonvisual Desktop Access (NVDA), and alternative input devices. The AWS Support App in Slack integration allows users to manage support cases directly within Slack without requiring separate console access. Amazon Q Developer provides AI-powered conversational assistance accessible from anywhere in the AWS Management Console, enabling users to create support cases and access help through natural language interactions.

AWS demonstrates strong commitment to accessibility through systematic evaluation against WCAG 2.1 and 2.2 Level AA standards, Section 508, and EN 301 549. AWS has received Accessibility Conformance Reports for over 160 AWS services, with comprehensive coverage across the service portfolio. These reports are available through AWS Artifact, providing customers with comprehensive accessibility compliance documentation.

AWS uses accessibility testing including automated tools to catch defects such as missing alternative text for graphics, missing form field labels, missing page titles, and incorrect semantic markup. Manual accessibility tests are performed with assistive technologies including screen readers, screen magnifiers, and voice recognition software to verify compatibility.

AWS gives our developer communities curated learning content to help them create accessible experiences for customers. We provide a powerful suite of in-person and on-demand, self-paced training sessions for our staff. These sessions range from introductory content and accessibility basics to more advanced accessibility tasks, such as how to use ARIA.

Skill Development and Training

AWS offers training and certification programs that build cloud fluency across organizations. The AWS Training and Certification portal offers learning paths tailored to specific roles—from cloud practitioners to specialized architects and developers—ensuring that each professional receives relevant education for their responsibilities. For more information, refer to https://aws.amazon.com/training/

Free digital training through AWS Skill Builder subscriptions provides an entry point for those new to AWS, with hundreds of on-demand courses covering foundational concepts and service-specific implementations. As skills advance, instructor-led training delivers deeper insights through hands-on labs and expert guidance. For organizations seeking to validate their teams' expertise, AWS Certification programs offer industry-recognized credentials that benchmark skills against established standards.

Beyond formal training, AWS enables continuous learning through resources like AWS Workshops, which provide guided tutorials for implementing specific solutions, and AWS Well-Architected Labs that teach best practices for secure, high-performing architectures. The annual re:Invent conference and regional summits create opportunities for in-person learning and networking with AWS experts and community members.

Technical Support

To address common technical challenges, AWS provides self-help resources such as troubleshooting guides that offer step-by-step solutions. This includes issues such as Amazon EC2 instance connectivity problems, AWS Lambda function errors, or Amazon S3 access denials.

When self-help resources aren't sufficient, support cases raised via AWS Support Center at https://console.aws.amazon.com/support connect customers directly with AWS engineers who are experts in specific services. The case management system allows for detailed documentation of issues, secure file sharing, and ongoing communication until resolution is achieved. For Business and Enterprise customers facing critical production issues, AWS's elevated support process engages senior engineers and service teams to drive rapid resolution.

AWS has also built proactive support mechanisms that identify and address potential issues before they impact customers. Systems like AWS Health Dashboard provide personalized alerts about events that might affect infrastructure, while service teams send notifications about important changes such as runtime deprecations or security patches.

Advanced Support Features

Customers can access AWS DevOps Agent (preview), an AI-powered frontier agent that resolves and proactively prevents incidents, working alongside teams to investigate issues and engage AWS Support with one click for faster resolution. Enterprise Support includes proactive services delivered by AWS experts, helping review cloud operations health, optimise costs, and scale workloads efficiently through workload reviews, best practices workshops, and deep dives.

Technical Account Managers (TAMs) lead Well-Architected Reviews, provide architectural guidance across security, reliability, performance efficiency, cost optimisation, and operational excellence, and bring in AWS subject matter experts when needed. TAMs execute critical events with confidence through AWS Countdown, optimize cloud costs through FinOps guidance, and build team expertise through TAM-led workshops and AWS GameDays.

6 Service Continuity & Exit

6.1.1 Business Continuity

Disaster Recovery Arrangements

AWS has designed its infrastructure and services with disaster recovery at their core, providing customers with multiple options to build resilient applications that can withstand various failure scenarios. The AWS global infrastructure is divided into Regions and Availability Zones, creating a foundation for effective disaster recovery strategies.

AWS offers several disaster recovery models that customers can implement based on their recovery time objectives (RTO) and recovery point objectives (RPO):

- **Backup & Restore**: The simplest approach where data is backed up to Amazon S3 and can be restored when needed. This approach is suitable for workloads with less stringent RTO requirements.
- **Pilot Light**: Core components of the system are always running in the cloud while the rest remain turned off until needed, allowing for faster recovery than backup-restore while minimizing costs.
- **Warm Standby**: A scaled-down version of a fully functional environment is always running, ready to scale up when disaster strikes, providing even faster recovery times.
- **Multi-Site Active/Active**: The full production environment is replicated across multiple AWS Regions, providing the fastest recovery but at higher cost.

AWS provides numerous services to support DR strategies, including AWS Backup for centralized backup management, Amazon S3 for durable object storage, and AWS Elastic Disaster Recovery (formerly CloudEndure Disaster Recovery) for rapid recovery of on-premises and cloud-based applications. For more details, refer to AWS Disaster Recovery: https://aws.amazon.com/disaster-recovery/

Service Resilience Measures

AWS builds resilience into its services through a multi-layered approach that combines infrastructure design, service architecture, and operational excellence:

Infrastructure Resilience

The AWS global infrastructure includes multiple geographically isolated Regions, each containing multiple Availability Zones (AZs). These AZs are physically separated data centers with independent power, cooling, and networking, but connected with high-bandwidth, low-latency networking. This design allows customers to architect applications that automatically fail over between AZs without interruption.

AWS also provides edge locations through Amazon CloudFront and AWS Global Accelerator that bring content delivery and network optimization closer to end users, improving performance and resilience.

Service Architecture

AWS services are designed with built-in redundancy and fault tolerance:

- **Amazon S3** provides 99.999999999% (11 nines) durability by automatically storing data redundantly across multiple devices and facilities.
- **Amazon RDS** offers multi-AZ deployments that automatically replicate database updates to a standby instance in a different AZ, with automatic failover during planned maintenance or instance failure.
- **Amazon DynamoDB** automatically replicates data across multiple AZs in a Region, with global tables providing multi-region replication.
- **Amazon Route 53** offers health checks and DNS failover to route traffic away from unhealthy endpoints.

Operational Excellence

AWS implements rigorous operational practices including:

- Continuous monitoring of service health
- Automated systems that detect and mitigate issues
- Regular game days to test resilience
- Post-incident analysis to drive continuous improvement

AWS also provides detailed Service Level Agreements (SLAs) for its services, documenting service availability commitments and service credit policies. For more details, refer to AWS Global Infrastructure at https://aws.amazon.com/about-aws/global-infrastructure/

Backwards Compatibility

AWS maintains a strong commitment to backwards compatibility, helping to ensure that customer applications and workloads continue to function as AWS services evolve. Our time-bound backwards compatibility provides flexibility for customers to plan and implement upcoming changes. This compatibility stops working beyond specified cut-off dates that are announced in advance. AWS maintains backwards compatibility by auto-assigning default values to fields required in new versions, ensuring existing integrations don't fail when new mandatory fields are introduced. For example, AWS maintains backward compatibility for legacy endpoints where applicable, such as Amazon S3's support for both path-style and virtual-hosted-style URL requests.

AWS's approach is guided by several key principles:

API Stability

AWS prioritizes API stability, meaning that once an API is released, AWS works to help ensure that applications built using that API continue to function even as the underlying

service evolves. When changes are necessary, AWS typically introduces new API versions rather than modifying existing ones, allowing customers to migrate at their own pace.

Deprecation Policies

When AWS must deprecate features or services, it follows a transparent process:

- **Advance Notice**: AWS provides significant advance notice before deprecating features, typically 12+ months for major changes.
- **Migration Paths**: AWS develops and documents clear migration paths to newer alternatives.
- **Transition Tools**: AWS often provides tools to assist with transitions, such as the EC2-Classic Network Migrator for moving from EC2-Classic to VPC.

Extended Support

AWS provides extended support for older versions of software with services like Amazon EKS, Amazon RDS, Amazon Aurora, and Amazon ElastiCache. Amazon EKS supports Kubernetes versions for 14 months under standard support, followed by an additional 12 months of extended support. Amazon RDS Extended Support allows customers to continue running databases on major engine versions past the end of standard support date for up to 3 years. This extended support provides critical security updates, bug fixes, and continued technical support for legacy versions, giving organizations additional time to plan and execute upgrades at their own pace.

Service Evolution

As AWS services evolve, we follow these practices:

- **Additive Changes**: New features are added without disrupting existing functionality.
- **Versioned Services**: For significant changes, AWS may introduce new versions (e.g., Lambda runtime versions) while continuing to support older versions.
- **Compatibility Layers**: When necessary, AWS provides compatibility layers to bridge old and new implementations.

Communication Channels

AWS communicates service changes through multiple channels:

- **AWS Personal Health Dashboard**: Provides alerts about changes that might affect your specific AWS resources.
- **AWS Service Health Dashboard**: Shows the general status of AWS services.
- **Release Notes and Documentation**: Details changes, new features, and migration guidance.
- **Deprecation Announcements**: Formal notifications about feature or service deprecations.
- **AWS Blog**: Provides detailed information about significant changes and best practices for adapting.

For more details, refer to AWS service health at https://status.aws.amazon.com/

Comprehensive Business Continuity

AWS's approach to business continuity extends beyond individual service resilience to provide customers with a complete toolkit for application availability:

- **AWS Well-Architected Framework**: Provides architectural best practices for designing reliable, secure, efficient, and cost-effective systems.
- **AWS Resilience Hub**: Helps customers assess and improve application resilience by defining resilience policies and validating that applications meet availability goals.
- **AWS Fault Injection Simulator**: Allows controlled chaos engineering experiments to test application resilience.
- **AWS Business Continuity Plan**: AWS maintains its own comprehensive business continuity plans, regularly tested through simulations and exercises.

6.1.2 By combining these capabilities with AWS's global infrastructure and service-specific resilience features, customers can build applications that maintain availability even during significant disruption events. For more details about AWS Resilience refer to: https://aws.amazon.com/resilience/Exit Planning (Lot 1a Specific)

Data Access Period

AWS provides a 90-day post-closure period for data access after contract termination. During this window, customers can access billing information, contact AWS Support, and reopen their account to retrieve remaining content. AWS retains all customer content for 30 days post-termination, allowing data retrieval without transfer fees (up to 100 GB monthly, with larger transfers eligible for fee waivers through Support approval).

Exit Procedures: Clear Decision Points and Buyer Involvement

AWS provides structured exit planning procedures with a 90-day post-closure period for data access, specific termination procedures, and pricing considerations during exit. Only the AWS account root user can close an AWS account, with account closure serving as notice of termination of the AWS Customer Agreement. The account closure process requires signing in as the root user to the account settings page, reading and understanding account closure guidance, and selecting Close Account in the dialog box. AWS provides a 90-day post-closure period during which customers can view past billing information and access AWS Support. Content and services that were not deleted or terminated before account closure remain on AWS hardware during this period.

For our enterprise customers, our Technical Account Managers can help facilitate structured exit planning with defined milestones and decision points throughout the process. These typically include initial notification, data migration planning, application transition sequencing, resource decommissioning schedules, and final account closure. For more information about AWS Enterprise Support, refer to https://aws.amazon.com/premiumsupport/plans/enterprise/

Terms During Exit: Pricing and Terms During Exit Period

AWS waives data transfer out charges for eligible customers when moving outside of AWS, available to all AWS customers globally and from any AWS Region. Customers moving up to 100 GB per month can transfer their data without fees. Customers transferring more than 100 GB need to contact AWS Support to review fee waiver requests, with AWS providing credits for migrated data upon approval.

Billing for on-demand charges stops during the post-closure period and is charged at the beginning of the next month. Customers remain responsible for outstanding fees and charges before closure. Subscription charges may continue after account closure. If customers reopen their account they might be charged for the cost of running AWS services during the post-closure period. AWS does not require minimum commitments or long-term contracts, though Reserved Instances offer optional minimum commitments. AWS provides

tools and services to help migrate from AWS, minimizing vendor lock-in. Upon termination of contract, AWS will not remove customer content from systems for 30 days following the termination date and will allow content retrieval if all amounts due under the customer agreement have been paid.

Our AWS Cost Explorer and AWS Budgets services remain available during exit to help you monitor and control spending as resources are decommissioned. For customers with specific exit pricing concerns, our sales teams can help incorporate customized terms regarding pricing stability, professional services assistance, and resource commitment management during your transition period. For more information about AWS pricing, refer to https://aws.amazon.com/pricing/

Exit Plan Provision: Timeline for Providing Exit Plan

The specific timeline requirements for an exit plan vary based on the contract type and customer requirements. Our enterprise engagement model supports developing formal exit plans according to your requirements to include specific terms in your AWS contracts. When timelines are specified contractually, our Enterprise teams work with you to deliver comprehensive exit plans that include current state assessment, migration strategies for different workloads, realistic timeframes, resource requirements, and risk mitigation approaches. We can update these plans throughout the contract lifecycle to reflect your evolving AWS usage and maintain alignment with current business continuity requirements.

AWS Professional Services provides comprehensive exit support services including:

- **Exit Strategy Development**. Working with customer teams to understand requirements, constraints, and timelines
- **Technical Assessment and Documentation**. Thorough assessments of the AWS environment
- **Data Migration Planning**. Expertise in planning and executing data migrations from AWS

For more information about AWS Professional Services, refer to https://aws.amazon.com/professional-services/

7 Innovation & Technology Advancement

AWS demonstrates a strong commitment to supporting innovation and emerging technologies through several key approaches:

Commitment to Innovation

A customer-driven development is central to AWS's approach—almost 90% of the AWS technology innovation is driven by customer feedback. We start by thinking deeply about customer goals and challenges to determine the services or solutions needed to solve them. We lead by delivering new services, then rapidly iterating upon those services based on customer feedback. In 2024, AWS released over 3,000 new significant services and features, in 2023 AWS released 3,410 new services and features, up from just 80 in 2011. This sustained growth reflects AWS's commitment to delivering products and services that matter to customers.

Access to New Capabilities and Features

AWS provides customers with immediate access to the latest technologies without requiring recapitalization of investments. We offer more than 240 cloud services spanning compute, storage, networking, databases, analytics, AI/ML, IoT, security, and emerging technologies

like quantum computing. We pioneered several industry-first innovations, including serverless computing with AWS Lambda (2014), Amazon SageMaker for machine learning, and ARM-based instances powered by AWS Graviton processors (2018). AWS Support provides immediate expertise in new services so customers can use them without waiting to build internal expertise.

AWS's innovation directly impacts customer business outcomes, with services such as Amazon Bedrock, Amazon Q, AWS Transform, and Amazon SageMaker allowing businesses to drive growth and generate value. Customers that migrate workloads to AWS can reduce total cost of ownership by up to 40% while gaining access to continuously improving capabilities.

Technology Roadmap Approach

AWS's roadmap approach is characterized by flexibility and customer responsiveness. We maintain "evergreen" customer agreements that allow AWS to remain nimble and cater to customers' ever-changing demands and requests. This constant pace of innovation—much of which comes from direct customer suggestions—is one of AWS's principal advantages.

We commit to providing at least 12 months' prior notice in the rare event AWS decides to discontinue a service or material functionality, helping customers plan for any required transitions.

Evergreen/Continuously Improving Ways of Working

AWS operates with a continuous improvement philosophy across multiple dimensions:

- **Operational excellence**: AWS continually improves existing services and frequently adds new services, helping customers maintain state-of-the-art IT infrastructure
- **Incident management**: Lessons learned from previous incidents inform improvements in response plans and workload architecture, driving a continuous improvement cycle to improve workload resiliency
- **Cost optimisation**: AWS has reduced prices 161 times since launching in 2006, passing efficiency gains and economies of scale to customers
- **Sustainability**: AWS relentlessly innovates its infrastructure design, build, and operations to make progress towards net-zero carbon by 2040 and being water positive by 2030

Additionally, our development, test, and production environments follow secure software development practices with security risk reviews prior to launch, allowing continuous improvement while maintaining security standards. Our established policies define roles, responsibilities, and classifications for managing changes to production environments, with development, test, and production environments logically separated to reduce risks of unauthorised access or change. This systematic approach to change management helps ensure that innovation can be delivered safely and reliably while maintaining operational excellence.

8 Standards Compliance & Sustainability

Environmental Initiatives and Data Centre Energy Efficiency

AWS operates with a global Power Usage Effectiveness (PUE) of 1.15 in 2024, significantly outperforming both the public cloud industry average of 1.25 and on-premises data centers at 1.63. AWS has implemented advanced data center designs and innovative cooling technologies to achieve this efficiency. AWS unveiled new data center components in 2024,

offering 12% more compute power with improved availability and efficiency. Additionally, All AWS services and data centers in EMEA, Singapore, and Indonesia are ISO 50001 certified for energy management.

AWS has implemented cooling technologies including direct-to-chip liquid cooling systems that reduce mechanical energy consumption by up to 46% without increasing water usage. In 2024, AWS achieved a global data center water usage effectiveness of 0.15 liters per kilowatt hour, representing a 17% improvement from 2023 and a 40% improvement since 2021. AWS is 53% of the way toward its water positive goal and has 23 global water replenishment projects that returned 4.3 billion liters of water to communities in 2024.

Net Zero Alignment Strategy

As part of The Climate Pledge, Amazon (including AWS) is committed to reaching net-zero carbon emissions by 2040—10 years ahead of the Paris Agreement. Over 590 companies have joined this pledge. Key achievements include:

- 100% of electricity consumed by Amazon was matched with renewable energy sources in 2024, for the second consecutive year, achieving the 2025 target five years early
- Over 600 global renewable energy projects across 29 countries
- Amazon's Climate Pledge Fund investment of $2 billion into clean technologies
- 4% reduction in carbon intensity in 2024 compared to 2023

AWS supports significant carbon footprint reductions for customers migrating to the cloud. According to a 2024 Accenture study, the AWS Cloud is up to 4.1 times more efficient than on-premises environments. AWS infrastructure can reduce carbon footprint by up to 99% for optimized workloads compared to on-premises environments.

Greening Government ICT Strategy Compliance

The UK government has established the Greening Government ICT and Digital Service Strategy 2020-2025, which includes extended mandatory reporting around carbon emissions, sustainability and social value. AWS provides tools and services that help UK government departments reduce their carbon footprint by migrating to the cloud.

- The Customer Carbon Footprint Tool uses simple visualizations to show customers their historical carbon emissions, estimate emissions avoided by using AWS instead of on-premises data centers, and review forecasted emissions based on their current use.
- The AWS Sustainability Data Fabric creates a unified data management system for environmental, social, and governance of information across organisations.
- The AWS Well-Architected Framework includes a Sustainability Pillar focusing on minimizing environmental impacts of cloud workloads.
- AWS Data Exchange makes it easy to find, subscribe to, and use third-party sustainability-related data in the cloud, including data sets made available via the Amazon Sustainability Data Initiative and the Open Data Sponsorship Program

Technology Code of Practice Alignment

AWS supports the UK Government's Technology Code of Practice (TCoP) through compliance capabilities, with over 6,500 government agencies already using AWS across all classification levels. AWS supports 143 security standards and compliance certifications, including the GDPR/Data Protection Act 2018, NHS Digital Data Security and Protection Toolkit, ISO 27001/27017/27018, Cyber Essentials Plus, and NCSC Cloud Security Principles. AWS provides TCO calculators and cost optimization tools that help government

departments understand and optimize their cloud spending. AWS supports open standards and interoperability, allowing government departments to avoid vendor lock-in and enabling data portability, as required by the TCoP

ISO 27001 Certification

AWS is certified under the ISO 27001 standard. ISO 27001 is a widely adopted global security standard that outlines the requirements for information security management systems. It provides a systematic approach to managing company and customer information that is based on periodic risk assessments. To achieve the certification, a company must show it has a systematic and ongoing approach to managing information security risks that affect the confidentiality, integrity, and availability of company and customer information.

AWS has established a formal programme to maintain the certification. More information regarding AWS's ISO 27001 certification can be found at http://aws.amazon.com/compliance/iso-27001-faqs/.

NCSC UK Cloud Security Principles

In 2016, National Cyber Security Centre (NCSC) UK published the Cloud Security Collection documents for public sector organisations that are considering the use of cloud services for handling information classified as OFFICIAL. The collection of guidance documents aims to help public sector organisations make informed decisions about cloud services and choose a cloud service that balances business benefits and security risks.

AWS is aligned with National Cyber Security Council Cloud Security Principles. Customers can find operational standard practices for these principles in the AWS Documentation portal here: https://docs.aws.amazon.com/config/latest/developerguide/operational-best-practices-for-ncsc.html.

We provide tools that allow customers control of access, movement, and security of their own data. These include:

- AWS Identity and Access Management (IAM), which allows customers to specify who or what can access services and resources in AWS, centrally manage fine-grained permissions, and analyse access to refine permissions across AWS
- AWS DataSync, the AWS Transfer Family and other services to move data
- A wide range of security tools and recommended practice guides to secure data.

The AWS Compliance Program helps customers to understand the controls in place at AWS to maintain security and compliance of the AWS Cloud. By tying together governance-focused, audit-friendly service features with applicable compliance or audit standards, AWS Compliance Enablers build on traditional programmes, helping customers to establish and operate in an AWS security control environment.

GDPR and processing of Personal Data

AWS offers a GDPR-compliant Data Processing Addendum (DPA), enabling customers to comply with GDPR contractual obligations. More information can be found at the following links:

- AWS GDPR Center: https://aws.amazon.com/compliance/gdpr-center/
- AWS Commitment to European Customer Data Protection: https://aws.amazon.com/compliance/eu-data-protection/
- Additional UK Addendum: https://d1.awsstatic.com/legal/aws-gdpr/UK_GDPR_Addendum_to_AWS_data_processing_addendum.pdf

Service Name

Amazon API Gateway

Service Overview

Amazon API Gateway is a fully managed service that enables developers to create, publish, maintain, monitor, and secure APIs at any scale. It serves as the front door for applications to access data, business logic, or functionality from backend services including AWS Lambda functions, EC2 instances, other AWS services, and HTTP endpoints. API Gateway supports RESTful APIs, HTTP APIs, and WebSocket APIs, handling critical tasks such as traffic management, authorization, throttling, monitoring, caching, and API version management while providing a unified access point for clients through public or private endpoints.

Key Use Cases

- API management and governance: Centralized platform for creating, deploying, and managing APIs with comprehensive lifecycle support
- Third-party integrations: Secure exposure of backend resources to external partners and developers with OAuth, throttling, and usage tiers
- Serverless applications: Integration with AWS Lambda functions to build serverless architectures with automatic scaling and pay-per-use pricing
- Microservices architecture: Acting as a unified access point for microservices deployed on Amazon ECS, EKS, or EC2 instances
- Mobile and web application backends: Providing HTTP endpoints for mobile apps and web applications with built-in CORS support
- IoT device communication: Enabling secure communication between IoT devices and backend services through WebSocket APIs
- Real-time applications: Supporting bidirectional communication for chat applications, real-time dashboards, and multiplayer games using WebSocket APIs

Features

- **Multiple API Types**: Support for REST APIs, HTTP APIs, and WebSocket APIs with different feature sets and pricing models
- **Integration Options**: Lambda proxy/custom integrations, HTTP proxy/custom integrations, AWS service integrations, mock integrations, and private integrations
- **Security and Authorization**: AWS IAM, Lambda authorizers, Amazon Cognito user pools, OAuth 2.0, OpenID Connect, API keys, and AWS WAF integration
- **Traffic Management**: Request throttling, usage plans, quotas, caching, and burst handling with token bucket algorithm

- **Endpoint Types**: Edge-optimized (CloudFront), Regional, and Private endpoints with VPC integration capabilities
- **Monitoring and Observability**: CloudWatch metrics, X-Ray tracing, access logging, and comprehensive API analytics
- **Developer Experience**: OpenAPI 3.0 support, SDK generation, developer portals, API documentation, and stage management
- **Data Transformation**: Request/response mapping, validation, binary media type support, and CORS configuration

Value Proposition

Amazon API Gateway offers compelling advantages as a fully managed service that eliminates the need for server provisioning and infrastructure management. It provides automatic scaling to handle any load without warm-up limitations, ensuring consistent performance from a few requests per day to thousands per second:

- The service offers comprehensive security features including native OAuth 2.0, IAM integration, and WAF protection, reducing development complexity
- API Gateway's pay-per-use pricing model with no upfront commitments makes it cost-effective, especially with the generous free tier offering one million API calls monthly
- The service integrates seamlessly with the broader AWS ecosystem, particularly Lambda for serverless architectures, while supporting hybrid deployments through private integrations and VPC connectivity

Service Integration

API Gateway integrates natively with numerous AWS services to enable comprehensive solutions. Primary integrations include AWS Lambda for serverless computing, Amazon EC2 and ECS for containerized workloads, and Application/Network Load Balancers for scalable backend services:

- Security integrations encompass AWS IAM for access control, Amazon Cognito for user authentication, and AWS WAF for API protection
- Monitoring capabilities leverage Amazon CloudWatch for metrics and logging, AWS X-Ray for distributed tracing, and AWS Config for compliance tracking
- Storage and database integrations include Amazon S3, DynamoDB, RDS, and other AWS services through direct service proxies
- Infrastructure management utilizes AWS CloudFormation for deployment automation, Route 53 for DNS management, and CloudFront for global content delivery through edge-optimized endpoints

Onboarding and Offboarding

Getting started with API Gateway involves creating an AWS account and accessing the service through the AWS Management Console, AWS CLI, or SDKs. The quickest path is using the HTTP API quick create feature, which automatically configures a Lambda integration, default route, and stage in minutes:

- Users can also import existing OpenAPI 3.0 definitions or build APIs incrementally using the console's guided workflow
- The service offers comprehensive tutorials for different integration types including Lambda functions, HTTP endpoints, and AWS services
- API Gateway provides CloudFormation templates for infrastructure as code deployments and supports automated CI/CD pipelines
- The generous free tier allows users to experiment with one million API calls monthly, making it risk-free to evaluate the service before production deployment

Getting Started Guide:
https://docs.aws.amazon.com/apigateway/latest/developerguide/getting-started.html

Data Removal

API Gateway offboarding involves deleting APIs, stages, deployments, and associated resources through the console, CLI, or APIs. Deleting an API removes all associated resources including methods, integrations, models, and documentation. Custom domain names, VPC links, and usage plans require separate deletion. API keys and client certificates are automatically cleaned up when no longer referenced. CloudWatch logs and X-Ray traces persist beyond API deletion and require separate cleanup. CloudFormation stacks can automate complete resource cleanup. The service provides immediate deletion for most resources, with some propagation delays for edge-optimized APIs due to CloudFront distribution updates.

Backup/Restore and Disaster Recovery

API Gateway automatically maintains service resilience through AWS's multi-AZ architecture but doesn't provide traditional backup capabilities for API configurations. Disaster recovery relies on API definition exports using OpenAPI specifications, CloudFormation templates, or SDK-based configuration exports. These can be version-controlled and stored in S3, Git repositories, or other systems for restoration. Cross-region deployment requires recreating APIs in target regions using exported definitions. The service supports blue-green deployments through stages and canary releases for safe updates. Regional failures require manual failover to pre-deployed APIs in other regions using Route 53 health checks.

Backup Capabilities

API Gateway does not integrate with AWS Backup service and lacks built-in backup functionality for API configurations. However, API definitions can be exported as OpenAPI specifications, CloudFormation templates, or programmatically through APIs for version control and restoration:

- These exports capture API structure, methods, integrations, models, and documentation but not runtime data like logs or metrics
- Users must implement custom backup strategies using these export mechanisms combined with infrastructure as code practices for comprehensive API configuration preservation and recovery capabilities

Disaster Recovery

API Gateway does not integrate with AWS Elastic Disaster Recovery service as it's a managed API proxy service without persistent storage requirements. DR strategies focus on multi-region API deployment using exported configurations and automated failover mechanisms:

- The service inherits AWS's regional resilience but requires manual intervention for cross-region failover
- Application Recovery Controller (ARC) can automate traffic routing between regions
- DR planning involves maintaining synchronized API definitions across regions, implementing health checks, and establishing automated deployment pipelines for rapid recovery

Recovery Objectives

API Gateway supports various RPO and RTO objectives through architectural design patterns. For low RTO requirements, multi-region active-active deployments with Route 53 health checks enable sub-minute failover. Warm standby approaches using pre-deployed APIs in secondary regions can achieve RTOs of 5-15 minutes. Cold backup and restore strategies using CloudFormation or OpenAPI exports typically require 30-60 minutes for full recovery. RPO is primarily determined by configuration change frequency since API Gateway doesn't store application data, with exports providing point-in-time configuration recovery capabilities.

Service Constraints

Service Quotas:
https://docs.aws.amazon.com/apigateway/latest/developerguide/limits.html

FAQ: https://aws.amazon.com/api-gateway/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/apigateway/latest/developerguide/welcome.html

User Guide: https://docs.aws.amazon.com/apigateway/latest/developerguide/getting-started.html

API Reference: https://docs.aws.amazon.com/apigateway/latest/api/

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/api-gateway/pricing/

Cost Optimization

API Gateway offers several cost optimization strategies starting with choosing HTTP APIs over REST APIs for simpler use cases, as HTTP APIs cost significantly less per million requests. Implementing response caching reduces backend calls and associated Lambda or compute costs:

- Request throttling and usage plans prevent unexpected cost spikes from traffic bursts
- Direct integration with AWS services eliminates intermediate Lambda functions for simple operations
- Regional endpoint selection over edge-optimized reduces CloudFront charges when serving localized traffic
- For private integrations, the new direct Application Load Balancer integration eliminates Network Load Balancer costs
- Efficient API design with fewer resources and consolidated endpoints reduces complexity and potential charges
- Monitoring usage patterns helps identify opportunities for caching, throttling adjustments, and architecture optimization

Savings Considerations

Primary cost savings come from API Gateway's serverless model eliminating server management overhead and capacity planning complexities. The pay-per-use pricing means customers only pay for actual API calls, avoiding fixed infrastructure costs:

- The generous free tier provides one million API calls monthly for 12 months, supporting development and small production workloads

- Caching capabilities can dramatically reduce backend compute costs by serving repeated requests from cache
- Direct AWS service integrations bypass Lambda costs for simple operations like S3 object retrieval or DynamoDB queries
- Efficient throttling prevents cost overruns during traffic spikes while maintaining service availability
- Consolidating multiple APIs reduces management overhead and potential duplicate charges
- Regional endpoint deployment eliminates unnecessary CloudFront costs for local traffic patterns, while still maintaining AWS's reliability and security benefits

Service Name

Amazon AppFlow

Service Overview

Amazon AppFlow is a fully-managed integration service that enables secure data exchange between software as a service (SaaS) applications and AWS services. It allows users to create data flows to transfer data between sources and destinations in minutes without writing code. The service supports data transformation, filtering, validation, and secure transfer with encryption both in transit and at rest. Users can run flows on demand, on event, or on schedule to keep data synchronized across applications while providing capabilities for data cataloging, partitioning, and custom connector development.

Key Use Cases

- Data Lake Hydration: Transfer data from SaaS applications like Salesforce, Marketo, ServiceNow, and Zendesk to Amazon S3 for analytics and machine learning at scale up to 100GB per flow
- Real-time Analytics: Create event-driven flows that transfer new records from Salesforce to Amazon Redshift for immediate analysis and reporting of sales opportunities
- Data Archiving and Historical Analysis: Schedule flows to transfer historical data from Google BigQuery or other sources to Amazon S3 for long-term storage and cost-effective analysis

Features

- **No-Code Data Integration**: Create data flows between SaaS applications and AWS services in minutes without writing code using an intuitive visual interface
- **Enterprise-Grade Data Transformations**: Map fields, concatenate data, mask sensitive information, truncate fields, and apply validations during data transfer
- **Flexible Trigger Mechanisms**: Run flows on-demand, on-event, or on-schedule with support for incremental and full data transfer modes
- **Data Privacy and Security**: Encrypt data in transit and at rest, support for AWS PrivateLink, customer-managed KMS keys, and IAM policy enforcement
- **Native SaaS Integrations**: Pre-built connectors for 60+ applications including Salesforce, Google Analytics, Slack, Zendesk, and many others
- **Custom Connector Development**: Build custom connectors using Python and Java SDKs for private APIs, on-premise systems, and other cloud services

Value Proposition

Amazon AppFlow provides significant advantages over building custom integrations or using other solutions through its fully managed infrastructure that eliminates the need for code development, server management, or scaling concerns. The service offers cost savings through pay-as-you-go pricing with no upfront fees, reducing both development time and operational overhead:

- Its enterprise-grade security features include automatic encryption, AWS PrivateLink support, and IAM integration, ensuring data protection without additional configuration
- The service provides high availability and reliability built on AWS infrastructure, with automatic retry mechanisms and error handling, while offering seamless integration with the broader AWS analytics and machine learning ecosystem for immediate data utilization

Service Integration

Amazon AppFlow integrates deeply with core AWS services to enable comprehensive data workflows. It natively connects with Amazon S3 for data lake storage, Amazon Redshift for data warehousing, and Amazon RDS for PostgreSQL for transactional data storage:

- The service integrates with AWS Glue Data Catalog for automatic data cataloging and schema discovery, enabling seamless analytics with Amazon Athena and other AWS analytics services
- AWS CloudTrail integration provides complete API audit trails, while AWS CloudFormation enables infrastructure-as-code deployment of flows and connections
- Amazon EventBridge integration allows event-driven architectures where AppFlow events can trigger other AWS services, and AWS Identity and Access Management (IAM) provides granular access control and security policies across all integrations

Onboarding and Offboarding

Getting started with Amazon AppFlow requires an AWS account and access credentials for source applications. Users begin by signing into the AWS Management Console and navigating to the Amazon AppFlow console:

- The setup process involves creating connections to source and destination systems using OAuth or API keys, then building flows through a visual interface that guides users through source selection, destination configuration, field mapping, and trigger setup
- Prerequisites include enabling APIs on third-party applications (like Google Cloud APIs for Google services), setting up necessary IAM roles for AWS service integrations, and optionally creating S3 buckets for data storage

- The entire setup can be completed in minutes without requiring code development or complex infrastructure provisioning

Getting Started Guide: https://docs.aws.amazon.com/appflow/latest/userguide/getting-started.html

Data Removal

Data removal during offboarding involves deleting Amazon AppFlow flows, connections, and associated AWS resources through the console or API. Users must delete flows first, then remove connector profiles and connections. Data stored in destination services like Amazon S3 buckets or Amazon Redshift tables requires separate cleanup actions. The service provides a DeleteFlow API operation that validates requests and can force deletion even for active flows. Complete offboarding also involves removing IAM roles, KMS keys, and any CloudFormation stacks created for AppFlow resources.

Backup/Restore and Disaster Recovery

Amazon AppFlow does not provide built-in backup and disaster recovery capabilities as a managed integration service. Data protection relies on the backup capabilities of destination services like Amazon S3 versioning, Amazon Redshift snapshots, or database backups. Flow configurations and metadata are managed by AWS but users should backup their connector configurations and flow definitions using AWS CloudFormation templates or API exports. Disaster recovery depends on the resilience of the underlying AWS infrastructure and the specific backup strategies implemented for destination data stores.

Backup Capabilities

Amazon AppFlow does not have native backup capabilities and does not integrate directly with AWS Backup service. The service itself is stateless for data processing, with flow configurations managed by AWS infrastructure:

- Data protection must be implemented at the destination level through services like Amazon S3 Cross-Region Replication, Amazon Redshift automated backups, or manual backup strategies for other destinations
- Flow configuration backup requires exporting flow definitions and recreating them if needed

Disaster Recovery

Amazon AppFlow does not integrate with AWS Elastic Disaster Recovery as it is a managed integration service without persistent infrastructure to protect. Disaster recovery for AppFlow workloads focuses on recreating flows and connections in alternate regions if needed:

- The service automatically handles infrastructure failures through AWS managed resilience
- Business continuity requires maintaining flow configuration documentation and ensuring destination services have appropriate disaster recovery strategies implemented

Recovery Objectives

Amazon AppFlow supports customer RPO and RTO objectives through its reliable, managed infrastructure and flexible scheduling options. The service provides near real-time data transfer capabilities with event-driven triggers that can minimize RPO for critical data flows. Multiple region availability enables geographic distribution of flows for improved RTO. However, specific RPO and RTO targets depend heavily on the backup and recovery capabilities of destination services rather than AppFlow itself, requiring customers to implement appropriate data protection strategies at destination endpoints.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/appflow/latest/userguide/service-quotas.html

FAQ: https://aws.amazon.com/appflow/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/appflow/latest/userguide/what-is-appflow.html

User Guide: https://docs.aws.amazon.com/appflow/latest/userguide/getting-started.html

API Reference: https://docs.aws.amazon.com/appflow/1.0/APIReference/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/appflow/pricing/

Cost Optimization

Amazon AppFlow offers several cost optimization opportunities through its pay-as-you-go pricing model with no upfront costs or infrastructure management fees. Users pay only for successful flow runs, avoiding charges for failed attempts or idle time:

- Cost optimization strategies include scheduling flows during off-peak hours to reduce API usage against source application limits, using incremental data transfer

modes to minimize data processing charges, and leveraging data filtering to transfer only necessary records

- The service eliminates the need for custom integration development and maintenance, significantly reducing total cost of ownership compared to building in-house solutions or managing third-party integration platforms

Savings Considerations

Primary cost savings come from eliminating the need to build and maintain custom integration code, reducing development time from weeks to minutes, and avoiding infrastructure provisioning and management costs. The service provides automatic scaling without capacity planning or server management fees:

- Additional savings result from reduced operational overhead through automated error handling, built-in security features, and managed updates
- Organizations can redirect engineering resources from integration maintenance to core business functionality
- The pay-per-use model ensures costs align directly with actual data transfer needs, avoiding fixed licensing fees or minimum commitments common with traditional integration platforms

Service Name

Amazon Athena

Service Overview

Amazon Athena is a serverless interactive query service that enables users to analyze data directly in Amazon S3 using standard SQL without requiring infrastructure setup or management. Users pay only for the queries they run, with no upfront costs or minimum fees. Athena automatically scales to execute queries in parallel, ensuring fast results even with large datasets and complex queries. The service supports various data formats including CSV, JSON, Parquet, ORC, and Avro, making it ideal for ad-hoc querying, data analysis, and business intelligence workloads. Athena can also run Apache Spark applications for more complex data processing tasks.

Key Use Cases

- Interactive ad-hoc SQL queries: Run quick queries on large datasets stored in Amazon S3 for troubleshooting performance issues and data exploration
- Log analysis and monitoring: Query AWS service logs, web logs, CloudTrail logs, and VPC flow logs for security analysis and operational insights
- Data lake analytics: Analyze staging data before loading into data warehouses and build interactive analytical solutions with notebook-based tools
- Business intelligence and reporting: Connect with visualization tools like Amazon QuickSight and other BI tools for creating dashboards and reports
- ETL operations: Use Create Table as Select (CTAS) statements for data transformation and federated queries for cross-source data integration
- Machine learning data preparation: Preprocess and analyze data for machine learning models using ML inference capabilities with Amazon SageMaker

Features

- **Serverless Architecture**: Zero infrastructure management with automatic scaling and pay-per-query pricing model
- **Standard SQL Support**: Full ANSI SQL compatibility with support for complex queries, joins, window functions, and user-defined functions
- **Multiple Data Format Support**: Native support for CSV, JSON, Parquet, ORC, Avro, and Apache Iceberg table formats
- **Federated Query**: Query data across multiple sources including S3, Redshift, DynamoDB, and external databases using connectors
- **Machine Learning Integration**: ML inference capabilities using Amazon SageMaker models directly within SQL queries

- **Apache Spark Support**: Run Apache Spark applications with simplified notebook experience for Python development
- **Workgroups and Query Management**: Organize users and queries with workgroups, query result management, and cost controls
- **Security and Compliance**: IAM integration, encryption at rest and in transit, VPC support, and fine-grained access controls

Value Proposition

Amazon Athena provides compelling advantages over traditional data warehousing and analytics solutions through its serverless architecture that eliminates infrastructure management overhead and reduces operational costs. The pay-per-query pricing model ensures customers only pay for actual usage without upfront investments or minimum commitments:

- Athena's direct integration with Amazon S3 enables immediate analysis of existing data without ETL processes or data movement
- The service scales automatically to handle varying workloads and provides consistent performance across different query complexities
- Native AWS service integration, robust security features, and support for standard SQL make it accessible to existing teams while federated query capabilities eliminate data silos by enabling cross-source analytics

Service Integration

Amazon Athena integrates deeply with core AWS services to provide a comprehensive analytics ecosystem. Amazon S3 serves as the primary data store, while AWS Glue Data Catalog provides metadata management and schema discovery:

- Amazon QuickSight enables data visualization and business intelligence reporting
- Integration with Amazon SageMaker allows ML model inference directly within queries
- AWS IAM provides access control and security management
- Additional integrations include CloudTrail for audit logs, CloudWatch for monitoring, Step Functions for workflow orchestration, and various data sources through federated query connectors
- Lake Formation provides fine-grained access controls, while services like Kinesis Data Firehose enable streaming data ingestion into S3 for Athena analysis

Onboarding and Offboarding

Getting started with Amazon Athena requires minimal setup and can be accomplished in minutes. Customers need an AWS account and an S3 bucket for storing query results:

- The process involves accessing the Athena console, creating a database, defining table schemas that point to data in S3, and running SQL queries
- AWS provides sample datasets and step-by-step tutorials to help new users understand the service quickly
- Athena automatically handles the underlying infrastructure provisioning and scaling
- For organizations, workgroups can be configured to manage user access, query limits, and cost controls
- The service integrates with existing AWS IAM policies for seamless security management and supports various client tools including JDBC/ODBC drivers

Getting Started Guide: https://docs.aws.amazon.com/athena/latest/ug/getting-started.html

Data Removal

Offboarding from Amazon Athena is straightforward since it's a serverless query service without persistent infrastructure. Customers need to delete their workgroups, named queries, and any custom data catalogs they created. Query results stored in S3 buckets must be manually deleted according to their retention policies. Table definitions in the AWS Glue Data Catalog can be removed, though the underlying data in S3 remains unchanged. Since Athena doesn't store customer data beyond query results, the primary consideration is managing S3 storage costs for historical query outputs and ensuring proper cleanup of metadata resources.

Backup/Restore and Disaster Recovery

Amazon Athena provides inherent disaster recovery through its serverless architecture and AWS's global infrastructure. The service automatically replicates across multiple Availability Zones within a region for high availability. Since Athena doesn't store customer data permanently, disaster recovery focuses on metadata and query definitions stored in AWS Glue Data Catalog, which supports cross-region replication. Query results stored in S3 benefit from S3's durability and can be replicated across regions. Workgroup configurations and named queries can be recreated using Infrastructure as Code approaches like CloudFormation for consistent disaster recovery.

Backup Capabilities

Athena itself doesn't require traditional backup mechanisms since it's a stateless query service. However, associated components need backup consideration:

- AWS Glue Data Catalog metadata can be backed up and replicated across regions
- Query results stored in S3 inherit S3's backup capabilities including versioning, cross-region replication, and integration with AWS Backup
- Named queries and workgroup configurations should be version-controlled and can be recreated using Infrastructure as Code

- The service doesn't integrate directly with AWS Backup since there's no persistent compute infrastructure to back up

Disaster Recovery

Amazon Athena supports disaster recovery through AWS's multi-AZ architecture and regional redundancy. The serverless nature means no persistent infrastructure requires recovery:

- Cross-region disaster recovery involves replicating AWS Glue Data Catalog metadata, S3 data, and recreating workgroups in alternate regions
- Athena doesn't integrate with AWS Elastic Disaster Recovery since there are no EC2 instances to replicate
- Recovery procedures focus on metadata restoration and ensuring data availability in the target region
- Applications can implement failover logic to redirect queries to alternate regions during outages

Recovery Objectives

Amazon Athena helps customers achieve aggressive RPO and RTO objectives through its serverless architecture and AWS's global infrastructure. The service provides near-zero RPO for metadata since AWS Glue Data Catalog changes can be replicated in real-time across regions. RTO depends primarily on the time to redirect queries to an alternate region and restore metadata, typically achievable within minutes. Since there's no infrastructure to recover, failover is primarily a matter of DNS redirection and ensuring data catalog availability in the recovery region.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/athena/latest/ug/service-limits.html

FAQ: https://aws.amazon.com/athena/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/athena/latest/ug/what-is.html

User Guide: https://docs.aws.amazon.com/athena/latest/ug/

API Reference: https://docs.aws.amazon.com/athena/latest/APIReference/

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/athena/pricing/

Cost Optimization

Amazon Athena offers unique cost optimization through its pay-per-query pricing model based on data scanned. Key strategies include using columnar formats like Parquet or ORC to reduce data scanned, implementing effective partitioning to limit scan scope, and leveraging compression to minimize storage costs:

- Query result reuse and caching can eliminate redundant processing
- Workgroups enable cost controls through query limits and monitoring
- Converting data to optimized formats using CTAS operations can significantly reduce long-term costs
- Using approximate functions and limiting result sets helps control query costs
- Managed query results feature eliminates separate S3 storage management overhead

Savings Considerations

Customers can achieve substantial cost savings by optimizing data storage formats, with Parquet providing up to 80% compression compared to CSV formats. Effective partitioning strategies can reduce data scanned by orders of magnitude:

- Using workgroups with cost controls prevents unexpected charges from runaway queries
- Leveraging Athena's serverless model eliminates infrastructure costs associated with traditional data warehouses
- Query optimization techniques like predicate pushdown and projection can significantly reduce processing costs
- Converting from traditional ETL processes to direct S3 querying eliminates data movement and storage duplication costs
- The lack of idle resource charges means customers only pay for active analysis time

Service Name

Amazon Aurora

Service Overview

Amazon Aurora is a fully managed relational database engine offered through Amazon RDS, compatible with MySQL and PostgreSQL. Aurora delivers up to 5x the performance of MySQL and 3x the performance of PostgreSQL without requiring application changes. Built for the cloud, Aurora features a distributed, fault-tolerant storage system that automatically grows to 128 TiB, automatic clustering and replication, and takes advantage of Amazon RDS management capabilities while providing enhanced performance and availability through innovative cloud-native architecture.

Key Use Cases

- Enterprise applications requiring high performance and availability with MySQL or PostgreSQL compatibility
- Web and mobile applications needing scalable database solutions with automatic scaling capabilities
- Data analytics and business intelligence workloads leveraging Aurora's performance optimization features
- Global applications requiring multi-region deployment using Aurora Global Database
- Variable workload applications using Aurora Serverless for cost-effective automatic scaling
- Gaming applications requiring consistent performance and scalability for player data and leaderboards

Features

- **High Performance**: Up to 5x MySQL and 3x PostgreSQL throughput with Aurora Parallel Query and Optimized Reads
- **Aurora Serverless**: Automatic scaling configuration that starts, stops, and scales based on application needs
- **Aurora Global Database**: Multi-region replication with low-latency global reads and disaster recovery
- **Storage Auto-scaling**: Distributed storage that automatically grows from 10 GiB to 128 TiB without downtime
- **Multi-AZ Availability**: High availability with Aurora Replicas and automatic failover across availability zones
- **Point-in-time Recovery**: Continuous incremental backups with restore capability to any second within 35 days

- **Zero-ETL Integration**: Direct integration with Amazon Redshift and SageMaker without complex data pipelines
- **Aurora Machine Learning**: Built-in ML capabilities with SageMaker and Amazon Comprehend integration

Value Proposition

Amazon Aurora provides superior price-performance compared to commercial databases with up to 90% cost reduction, while delivering 5x MySQL and 3x PostgreSQL performance. Aurora eliminates database administration overhead with fully managed operations, automatic scaling, and built-in high availability:

- The service offers seamless MySQL and PostgreSQL compatibility, enabling easy migration without application changes
- Aurora's cloud-native architecture provides innovative features like Aurora Serverless for variable workloads, Global Database for worldwide applications, and zero-ETL integrations that traditional databases cannot match

Service Integration

Aurora integrates deeply with Amazon RDS for management capabilities, Amazon S3 for data import/export and backup storage, AWS Lambda for serverless compute integration, Amazon CloudWatch for monitoring and logging, and AWS Backup for centralized backup management. Aurora also connects with Amazon Redshift through zero-ETL integration for analytics, Amazon SageMaker for machine learning workloads, AWS DMS for database migration, Amazon VPC for network isolation, and IAM for access control. Additional integrations include Application Auto Scaling for read replica scaling and Amazon DevOps Guru for performance insights.

Onboarding and Offboarding

Customers get started by completing environment setup tasks, then creating an Aurora DB cluster through the Amazon RDS console by selecting Aurora MySQL or Aurora PostgreSQL as the engine. The process involves choosing instance classes, configuring networking with VPC and security groups, setting up database credentials, and configuring backup and monitoring options:

- Aurora offers push-button migration tools from existing RDS MySQL and PostgreSQL databases through snapshots or replication
- Quick start templates and tutorials guide users through creating their first cluster and connecting applications

Getting Started Guide:
https://docs.aws.amazon.com/AmazonRDS/latest/AuroraUserGuide/CHAP_GettingStartedAurora.html

Data Removal

Customers can remove data by deleting Aurora DB clusters, which automatically deletes associated instances and storage. Manual snapshots must be explicitly deleted as they persist beyond cluster deletion. Automated backups are automatically removed when clusters are deleted. For complete data removal, customers should delete all snapshots, disable automated backups, and ensure any exported data in Amazon S3 is also removed from their account.

Backup/Restore and Disaster Recovery

Aurora provides automated continuous incremental backups with point-in-time recovery to any second within a 1-35 day retention period. Manual snapshots can be created and shared across accounts. Aurora integrates with AWS Backup for centralized backup management, cross-region copying, and compliance features like Vault Lock. Aurora Global Database enables disaster recovery across regions with RPO near zero and RTO of less than 1 minute.

Backup Capabilities

Aurora automatically performs continuous incremental backups to Amazon S3 without performance impact. Manual database snapshots can be created on-demand and retained indefinitely:

- Aurora integrates with AWS Backup service for centralized backup policies, cross-region replication, and compliance features
- Backup storage is billed separately and provides point-in-time recovery capabilities within the configured retention period

Disaster Recovery

Aurora supports disaster recovery through Aurora Global Database with cross-region replication, automated failover capabilities, and read replicas that can be promoted to primary. Aurora Replicas provide within-region DR with automatic failover:

- Cross-region snapshots enable recovery in different regions
- Aurora does not directly integrate with AWS Elastic Disaster Recovery but provides comprehensive DR through native capabilities

Recovery Objectives

Aurora helps achieve aggressive RPO and RTO objectives through continuous incremental backups enabling RPO as low as seconds, and Aurora Global Database providing RPO near zero across regions. RTO can be achieved in minutes through automated failover to Aurora Replicas or Global Database secondary regions. Point-in-time recovery and automated backup features ensure data protection while Aurora's high availability design minimizes downtime scenarios.

Service Constraints

Service Quotas:
https://docs.aws.amazon.com/AmazonRDS/latest/AuroraUserGuide/CHAP_Limits.html

FAQ: https://aws.amazon.com/rds/aurora/faqs/

Technical Requirements

Service Documentation:
https://docs.aws.amazon.com/AmazonRDS/latest/AuroraUserGuide/

User Guide:
https://docs.aws.amazon.com/AmazonRDS/latest/AuroraUserGuide/CHAP_GettingStartedAurora.html

API Reference:
https://docs.aws.amazon.com/AmazonRDS/latest/AuroraUserGuide/ProgrammingGuide.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/rds/aurora/pricing/

Cost Optimization

Aurora offers several unique cost optimization options including Aurora Serverless v2 for automatic scaling based on demand, eliminating idle capacity costs. Aurora I/O-Optimized configuration reduces I/O costs for high-throughput workloads:

- Reserved Instances provide significant discounts for predictable workloads
- Storage auto-scaling ensures you only pay for used storage
- Aurora's shared storage architecture reduces costs compared to traditional replication by eliminating duplicate storage for replicas
- Database cloning creates cost-effective copies for testing without duplicating storage

Savings Considerations

Primary cost savings come from Aurora's innovative architecture eliminating the need for separate storage for replicas, reducing infrastructure costs by up to 50%. Aurora Serverless automatically scales to zero during idle periods, eliminating costs for unused capacity:

- The service reduces operational costs through fully managed operations, automated patching, and built-in monitoring
- Migration from commercial databases can achieve up to 90% cost reduction
- Aurora's performance improvements can reduce the number of required instances
- Zero-ETL integrations eliminate separate ETL infrastructure costs for analytics workloads

Service Name

Amazon Bedrock

Service Overview

Amazon Bedrock is a fully managed service that provides access to high-performing foundation models from leading AI companies and Amazon through a unified API. It enables developers to build generative AI applications with comprehensive capabilities including security, privacy, and responsible AI features. The service offers foundation models from providers like Amazon, Anthropic, AI21 Labs, Cohere, Meta, Mistral AI, OpenAI, and others. Users can experiment with various models, customize them privately with their data using techniques like fine-tuning and retrieval-augmented generation (RAG), and create managed agents that execute complex business tasks without managing infrastructure.

Key Use Cases

- Content generation and document processing for marketing materials, reports, and creative writing
- Conversational AI and chatbots for customer service and support applications
- Code generation, optimization, and vulnerability scanning for software development
- Data analysis and insights extraction from documents, images, videos, and audio files
- Automated business process workflows including travel booking, insurance claims processing, and inventory management
- Personalized recommendations and customer engagement across various industries

Features

- **Foundation Model Choice**: Access to industry-leading foundation models from multiple providers through a single API
- **Model Customization**: Fine-tuning, continued pretraining, and model distillation capabilities for domain-specific adaptation
- **Knowledge Bases and RAG**: Retrieval-augmented generation with support for multimodal data and structured data retrieval
- **Agents and Multi-Agent Collaboration**: Managed agents that execute complex tasks and coordinate with each other for sophisticated workflows
- **Guardrails**: Built-in safeguards against harmful content with automated reasoning and contextual grounding checks

- **Data Automation**: Automated processing and insight extraction from documents, images, videos, and audio files
- **Model Evaluation**: Automatic and human-based evaluation capabilities with LLM-as-a-Judge functionality
- **Service Tiers**: Priority, Standard, and Flex tiers to optimize performance and cost for different workload types

Value Proposition

Amazon Bedrock eliminates the complexity of managing AI infrastructure while providing enterprise-grade security, privacy, and compliance. The service offers unparalleled model choice from leading AI providers through a single API, reducing vendor lock-in and enabling easy model switching:

- Built-in capabilities like Guardrails ensure responsible AI deployment, while features like Knowledge Bases and Agents enable sophisticated applications without extensive development effort
- The serverless architecture means no infrastructure management, and flexible pricing options including service tiers allow cost optimization
- Integration with AWS services provides seamless deployment within existing cloud architectures

Service Integration

Amazon Bedrock integrates deeply with core AWS services including Amazon S3 for data storage and Knowledge Bases, AWS Lambda for serverless functions and agent actions, Amazon DynamoDB for state management, Amazon VPC for secure networking, AWS IAM for access control and permissions, Amazon CloudWatch for monitoring and logging, AWS CloudTrail for audit trails, Amazon EventBridge for event-driven architectures, Amazon OpenSearch for vector databases in Knowledge Bases, and AWS Step Functions for orchestrating complex workflows. The service also works with Amazon SageMaker for model training and evaluation, AWS Backup for data protection, and integrates into Amazon SageMaker Unified Studio for collaborative development environments.

Onboarding and Offboarding

Getting started requires an AWS account with proper IAM permissions for Amazon Bedrock. Beginning June 15, 2025, access to foundation models is enabled by default, though first-time Anthropic model users may need to submit use case details:

- Users can start through the Amazon Bedrock console playgrounds for text and image generation, access the service via APIs and SDKs, or follow step-by-step tutorials

- The setup involves creating IAM roles with necessary policies, requesting model access if needed, and choosing appropriate foundation models for specific use cases
- Multiple getting started resources including tutorials, workshops, and code examples are available to accelerate onboarding

Getting Started Guide: https://docs.aws.amazon.com/bedrock/latest/userguide/getting-started.html

Data Removal

Amazon Bedrock follows AWS data handling practices where customer data is not retained after processing. Users can delete custom models, Knowledge Bases, agents, and associated resources through the console or APIs. Data stored in integrated services like S3 must be deleted separately by the customer. The service provides secure data handling with no sharing of customer inputs or outputs with third-party model providers, ensuring complete data privacy during and after service usage.

Backup/Restore and Disaster Recovery

Amazon Bedrock is a managed service where AWS handles infrastructure resilience and availability across multiple Availability Zones. Customer-created resources like custom models, Knowledge Bases, and agent configurations are maintained by AWS with built-in redundancy. For customer data in integrated services like S3, standard AWS backup and disaster recovery capabilities apply. The service itself does not require customer-managed backup procedures as the foundation models and service infrastructure are managed by AWS.

Backup Capabilities

Amazon Bedrock does not require traditional backup capabilities as it provides access to foundation models and managed services. Custom models created through fine-tuning are stored and maintained by AWS:

- Customer data in Knowledge Bases relies on the underlying data sources in S3 or other supported services
- AWS Backup integration is available for the underlying storage services that customers use with Bedrock, but the service itself operates as a managed offering without customer backup requirements

Disaster Recovery

Amazon Bedrock provides built-in disaster recovery through AWS's multi-AZ architecture and global infrastructure. The service is designed for high availability with automatic failover capabilities:

- While the service does not directly integrate with AWS Elastic Disaster Recovery, the underlying infrastructure follows AWS disaster recovery best practices
- Customer applications using Bedrock should implement their own disaster recovery strategies for application-level components and data stored in integrated AWS services

Recovery Objectives

Amazon Bedrock supports customer RPO and RTO objectives through its highly available, managed architecture with automatic scaling and multi-AZ deployment. The service provides consistent API availability and can handle traffic spikes without manual intervention. For custom models and configurations, AWS maintains redundancy to ensure quick recovery. Customers can achieve low RTO by leveraging multiple AWS regions and implementing proper error handling and retry logic in their applications using Bedrock APIs.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/bedrock/latest/userguide/quotas.html

FAQ: https://aws.amazon.com/bedrock/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/bedrock/latest/userguide/what-is-bedrock.html

User Guide: https://docs.aws.amazon.com/bedrock/latest/userguide/getting-started.html

API Reference:
https://docs.aws.amazon.com/bedrock/latest/APIReference/welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/bedrock/pricing/

Cost Optimization

Amazon Bedrock offers unique cost optimization through service tiers (Priority, Standard, Flex) allowing workload-specific pricing, with Flex tier providing significant discounts for non-time-sensitive tasks. Prompt caching reduces costs up to 90% by caching repeated prompt prefixes:

- Model Distillation creates smaller, faster models that are up to 75% less expensive with minimal accuracy loss

- Intelligent Prompt Routing automatically selects the most cost-effective model, reducing costs up to 30%
- Batch processing provides cost advantages for large-scale workloads, while Provisioned Throughput offers predictable pricing for consistent workloads
- The pay-per-use model with token-based pricing ensures customers only pay for actual usage

Savings Considerations

Main cost savings come from the serverless architecture eliminating infrastructure costs and the ability to choose optimal models for each task rather than over-provisioning. Service tier selection allows matching cost to performance requirements, with Flex tier offering substantial savings for batch workloads:

- Prompt engineering and management reduce token usage through optimized prompts
- Model customization reduces the need for expensive fine-tuning by leveraging Knowledge Bases and RAG
- Multi-model access prevents vendor lock-in enabling cost-effective model switching
- The elimination of model hosting, scaling, and maintenance costs provides significant operational savings compared to self-managed AI infrastructure

Service Name

Amazon Braket

Service Overview

Amazon Braket is a fully managed AWS service that provides researchers, scientists, and developers with access to quantum computing technologies. It offers a single point of access to various quantum computers from third-party providers including AQT, IonQ, IQM, Rigetti, and QuEra, along with quantum circuit simulators. The service enables users to explore and design quantum and hybrid algorithms, test algorithms on simulators, and run quantum programs on different types of quantum hardware including superconducting, trapped-ion, and neutral-atom quantum computers. Braket includes development environments with pre-installed Jupyter notebooks, the Amazon Braket SDK, and integration with frameworks like PennyLane for variational quantum algorithms in the current Noisy Intermediate-Scale Quantum (NISQ) era.

Key Use Cases

- Quantum algorithm research and development: Design, test, and optimize quantum algorithms for computational problems beyond classical computer reach
- Hybrid quantum-classical computing: Develop variational quantum algorithms combining classical and quantum resources using PennyLane framework
- Quantum simulation and modeling: Simulate quantum systems and molecular behavior for drug discovery, materials science, and chemical engineering applications
- Optimization problems: Solve complex optimization challenges in financial portfolio management, logistics, and machine learning using quantum computing advantages
- Educational and training purposes: Learn quantum computing fundamentals through managed Jupyter notebooks with pre-built examples and tutorials

Features

- **Multi-Provider Quantum Hardware Access**: On-demand access to quantum computers from AQT, IonQ, IQM, Rigetti, and QuEra with different qubit technologies
- **Quantum Circuit Simulators**: Four simulator options including local simulator, SV1, DM1, and TN1 for testing and validating circuits
- **Braket SDK and Development Tools**: Hardware-agnostic SDK for designing quantum algorithms with support for multiple programming languages and frameworks

- **Hybrid Jobs**: Managed execution environment for quantum-classical hybrid algorithms with embedded simulators
- **Braket Direct**: Dedicated device access through reservations with expert advice and experimental capabilities
- **Jupyter Notebook Environment**: Fully managed notebooks pre-installed with sample algorithms, resources, and developer tools

Value Proposition

Amazon Braket eliminates the traditional barriers to quantum computing by providing cost-effective, on-demand access to multiple quantum hardware providers without upfront commitments or individual provider negotiations. Unlike building quantum computing capabilities in-house, Braket offers a fully managed service that handles underlying infrastructure complexity, including high-performance simulators running on Amazon EC2 clusters:

- The service provides unique advantages through its hardware-agnostic approach, allowing algorithm testing across different quantum technologies with a single line of code change
- raket's integration with S services enables seamless scaling, monitoring, and data management, while the pay-as-you-go pricing model makes quantum computing accessible to organizations of all sizes, from research institutions to enterprises exploring quantum advantages

Service Integration

Amazon Braket integrates deeply with core AWS services for comprehensive quantum computing workflows. Amazon S3 serves as the primary storage for quantum computation results and data management:

- Amazon SageMaker provides the underlying infrastructure for managed Jupyter notebook instances and machine learning integration
- AWS Identity and Access Management (IAM) controls access permissions and security policies for quantum resources
- Amazon CloudWatch monitors Braket resources and provides metrics for quantum tasks
- AWS CloudTrail logs all API calls for audit and compliance purposes
- Amazon EventBridge receives quantum computing events including task state changes and device status updates
- Amazon EC2 powers the high-performance classical computing clusters used for quantum circuit simulation, while AWS Lambda can be used for serverless quantum computing workflows

Onboarding and Offboarding

Customers begin by enabling Amazon Braket in their AWS account and creating a managed Jupyter notebook instance through the AWS Management Console. The service provides pre-installed example notebooks, tutorials, and the Amazon Braket SDK to accelerate learning:

- Users can start with the local quantum simulator for development, progress to on-demand simulators (SV1, DM1, TN1), and then access real quantum hardware when ready
- The getting started process includes building a first Bell state quantum circuit, running it on simulators, and optionally executing on quantum processors
- AWS offers a free tier with one hour of simulation time monthly for the first year
- Educational institutions can apply for AWS research credits, while enterprise customers can leverage the Amazon Quantum Solutions Lab for expert guidance and collaborative development of quantum applications

Getting Started Guide:
https://docs.aws.amazon.com/braket/latest/developerguide/braket-get-started.html

Data Removal

Amazon Braket stores quantum computation results in customer-owned Amazon S3 buckets, giving users full control over their data lifecycle and deletion. When offboarding, customers can delete their S3 buckets and associated data following standard S3 data removal procedures. Notebook instances can be stopped and deleted through the Braket console, removing associated local storage. Users should terminate any running hybrid jobs and delete created IAM roles and policies. CloudWatch logs and CloudTrail records follow their respective retention policies and can be manually deleted if needed.

Backup/Restore and Disaster Recovery

Amazon Braket does not provide traditional backup and disaster recovery capabilities as it primarily serves as a compute service for quantum algorithms rather than a data storage service. Quantum computation results are stored in customer-controlled Amazon S3 buckets, which inherit S ’s durability and availability features. sers are responsible for implementing backup strategies for their quantum algorithms, notebooks, and results using standard AWS backup services. The service itself is designed for stateless quantum task execution, where reproducibility comes from re-running quantum circuits rather than restoring previous states.

Backup Capabilities

Amazon Braket does not have native backup capabilities as it functions as a quantum computing execution service rather than a persistent data storage service. The service stores computation results in customer-owned S3 buckets, which support versioning,

cross-region replication, and integration with AWS Backup for comprehensive data protection. Customers must implement their own backup strategies for quantum algorithms, notebook code, and research data using standard AWS backup services and S3 features.

Disaster Recovery

Amazon Braket does not directly integrate with AWS Elastic Disaster Recovery as it provides stateless quantum computing execution rather than persistent infrastructure that requires disaster recovery. The service operates across multiple AWS regions (US East, US West, Europe) providing geographic distribution. Quantum algorithms and code can be replicated across regions, and results stored in S benefit from S 's cross-region replication capabilities for disaster recovery scenarios.

Recovery Objectives

Amazon Braket supports recovery objectives through its multi-region availability and stateless execution model. Since quantum computations are reproducible by re-running circuits, Recovery Time Objective (RTO) is primarily limited by quantum hardware queue times and circuit execution duration. Recovery Point Objective (RPO) depends on customer backup strategies for quantum algorithms and result data stored in S3. The service's integration with S cross-region replication enables near-zero RPO for critical quantum computation results and algorithms.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/braket/latest/developerguide/braket-quotas.html

FAQ: https://aws.amazon.com/braket/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/braket/latest/developerguide/what-is-braket.html

User Guide: https://docs.aws.amazon.com/braket/latest/developerguide/what-is-braket.html

API Reference: https://docs.aws.amazon.com/braket/latest/APIReference/index.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/braket/pricing/

Cost Optimization

Amazon Braket offers unique cost optimization through its pay-as-you-go model with no upfront commitments and spending limits for quantum processing units (QPUs) to control costs proactively. The service provides a free local simulator and one hour of monthly simulation time in the AWS Free Tier:

- Cost optimization strategies include testing circuits on free simulators before using paid QPUs, leveraging the near real-time cost tracking in the Braket SDK, setting per-device spending limits, and using TN1 simulator with low shot counts for initial testing
- Braket Direct offers reserved device access at hourly rates rather than per-task pricing for high-volume users
- Educational institutions can access AWS research credits, and the service provides cost estimation tools within development notebooks

Savings Considerations

Primary cost savings come from eliminating the need for expensive quantum computing hardware investments and maintenance costs. The shared access model reduces individual organization costs compared to dedicated quantum computer procurement:

- Pay-per-use pricing eliminates capacity planning risks and unused resource costs
- Free simulation capabilities enable algorithm development and testing without hardware costs
- The managed service approach reduces staffing requirements for quantum computing infrastructure expertise
- Integration with AWS services leverages existing cloud investments and skills
- Educational and research programs provide significant cost reductions through credits and grants
- The ability to test across multiple quantum hardware types with minimal switching costs enables optimal price-performance selection for specific applications

Service Name

Amazon CloudFront

Service Overview

Amazon CloudFront is a web service that speeds up distribution of static and dynamic web content to users through a worldwide network of data centers called edge locations. When a user requests content, CloudFront routes the request to the edge location that provides the lowest latency, delivering content with the best possible performance. If content is already cached at the edge location, CloudFront delivers it immediately; if not, CloudFront retrieves it from the origin server. CloudFront accelerates website performance by reducing the load on application servers and delivering content from locations closest to end users, utilizing the AWS backbone network for optimal performance.

Key Use Cases

- Accelerating static website content delivery: CloudFront speeds up delivery of static content like images, style sheets, and JavaScript by using Amazon S3 as the origin
- Video on demand (VOD) and live streaming: CloudFront supports streaming in common formats and caches media fragments at the edge to reduce load on origin servers
- API and application acceleration: CloudFront can accelerate dynamic content delivery and API responses through intelligent routing and protocol optimizations
- Global content distribution: Delivering websites, applications, and APIs to customers globally with low latency and high transfer speeds
- Secure content delivery: Serving private content using signed URLs, signed cookies, field-level encryption, and integration with security services

Features

- **Global Edge Network**: Worldwide network of edge locations including Points of Presence (POPs), Regional Edge Caches (RECs), and embedded Points of Presence for content delivery
- **Origin Shield**: Additional caching layer that reduces traffic to origin servers and improves cache hit ratios while providing centralized caching
- **CloudFront Functions**: Lightweight JavaScript functions that run at edge locations for high-scale, latency-sensitive CDN customizations with sub-millisecond execution
- **Lambda@Edge**: Serverless compute feature for complex, computationally intensive operations that run closer to viewers in Node.js or Python

- **Security Integration**: Built-in protection through AWS Shield, AWS WAF, field-level encryption, SSL/TLS support, and origin access control
- **Real-time Monitoring**: Integration with CloudWatch for metrics, standard and real-time logging capabilities, and comprehensive request tracking
- **Multiple Origin Support**: Support for various origins including S3, MediaPackage, Application Load Balancers, EC2 instances, and custom HTTP servers with failover capabilities

Value Proposition

CloudFront delivers superior global performance through its extensive edge network with hundreds of terabits of deployed capacity and connections to thousands of telecom carriers. The service provides automatic DDoS protection via AWS Shield, integrated security through WAF, and comprehensive compliance certifications including PCI-DSS Level 1, HIPAA, and SOC:

- CloudFront offers significant cost advantages with free data transfer between AWS services, regional edge caches that reduce origin costs, and flexible pricing options including flat-rate plans with no overages
- The platform enables advanced customization through edge functions while maintaining enterprise-grade reliability and fast change propagation within minutes

Service Integration

CloudFront integrates seamlessly with Amazon S3 for static content storage and origin access control, AWS Shield and AWS WAF for comprehensive security protection, and Amazon Route 53 for DNS management and traffic routing. The service works with AWS Certificate Manager for SSL/TLS certificates, Amazon CloudWatch for monitoring and logging, and AWS Lambda for edge computing capabilities:

- CloudFront also integrates with Amazon API Gateway for API acceleration, Application Load Balancers and Network Load Balancers for dynamic content, AWS Elemental MediaPackage and MediaStore for video streaming, and Amazon ElastiCache for enhanced caching strategies
- Additional integrations include AWS Config for compliance, AWS CloudTrail for audit logging, and AWS Resource Access Manager for resource sharing

Onboarding and Offboarding

Getting started with CloudFront involves setting up an AWS account, then creating a distribution by specifying origin servers where the original content resides. Customers can use the AWS Management Console to create a standard distribution with S3 as origin, which automatically configures origin access control (OAC) for secure authenticated requests:

- The process includes uploading content to the origin, creating the CloudFront distribution, and configuring cache behaviors and security settings
- AWS provides multiple tutorials including console-based setup, AWS CLI implementation, and secure static website deployment
- The service offers standard distributions for typical use cases and multi-tenant distributions for SaaS providers managing multiple tenants

Getting Started Guide: https://docs.aws.amazon.com/AmazonCloudFront/latest/DeveloperGuide/GettingStarted.html

Data Removal

To offboard from CloudFront, customers must first disable their distributions, then delete them after confirming deployment status. The process involves updating distribution configuration to disabled status, waiting for the change to propagate across edge locations, then submitting a delete request with the distribution's ETag. Customers should also clean up associated resources including S3 buckets, SSL certificates, CloudFront Functions, Lambda@Edge functions, and any created IAM roles or policies to ensure complete data removal and avoid ongoing charges.

Backup/Restore and Disaster Recovery

CloudFront inherently provides disaster recovery capabilities through its globally distributed architecture with multiple edge locations and Regional Edge Caches. The service offers origin failover functionality that automatically routes traffic to backup origins when primary origins become unavailable. CloudFront's multi-origin support enables backend architecture redundancy, while Origin Shield provides additional protection and centralized caching. The service maintains content availability through cached objects across hundreds of edge locations worldwide, ensuring content remains accessible even during regional outages.

Backup Capabilities

CloudFront does not provide traditional backup capabilities as it is a content delivery network focused on caching and distribution rather than data storage. Content backup is managed at the origin level (such as S3 versioning, database backups, or application server backups). CloudFront automatically caches content across its global network of edge locations, providing natural redundancy and availability, but this caching mechanism is designed for performance optimization rather than data protection or backup purposes.

Disaster Recovery

CloudFront supports disaster recovery through its distributed architecture and origin failover capabilities. The service does not directly integrate with AWS Elastic Disaster Recovery, but provides origin failover that automatically switches to backup origins during

primary origin failures. CloudFront's global network of edge locations ensures content remains available even during regional disasters, and customers can configure multiple origins across different AWS regions for comprehensive disaster recovery strategies.

Recovery Objectives

CloudFront helps meet aggressive RPO and RTO objectives through its global edge network that maintains cached content across hundreds of locations worldwide. Origin failover provides near-instantaneous switching to backup origins, supporting RTO objectives of seconds to minutes. The service's distributed caching architecture ensures content remains available from edge locations even during origin failures, effectively providing RPO of near-zero for cached content. Multi-region origin configurations and intelligent traffic routing further enhance recovery capabilities for mission-critical applications.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/AmazonCloudFront/latest/DeveloperGuide/cloudfront-limits.html

FAQ: https://aws.amazon.com/cloudfront/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/AmazonCloudFront/latest/DeveloperGuide/

User Guide: https://docs.aws.amazon.com/AmazonCloudFront/latest/DeveloperGuide/

API Reference: https://docs.aws.amazon.com/cloudfront/latest/APIReference/

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/cloudfront/pricing/

Cost Optimization

CloudFront offers unique cost optimization through flat-rate pricing plans (Free, Pro, Business, Premium) that eliminate overage charges and provide predictable monthly costs with included AWS WAF and DDoS protection. Price class selection allows customers to limit content delivery to specific geographic regions, reducing costs by excluding more expensive edge locations:

- Origin Shield reduces origin operating costs by collapsing requests and improving cache hit ratios

- Free data transfer between AWS services and CloudFront for origin fetches provides significant savings
- Regional Edge Caches decrease operational burden on origins while custom pricing options are available for high-volume customers with 10TB+ monthly commitments

Savings Considerations

Primary cost savings come from reduced origin server load through efficient caching and Origin Shield implementation, which can significantly decrease compute and bandwidth costs at origin servers. CloudFront's free tier includes 1TB of data transfer and 10 million requests monthly, providing substantial value for smaller workloads:

- Intelligent caching strategies and proper cache policy configuration maximize cache hit ratios, reducing expensive origin requests
- Geographic optimization through price class selection and strategic content placement near user populations minimizes data transfer costs
- Integration with other AWS services provides additional cost efficiencies through consolidated billing and service-specific optimizations like S3 Transfer Acceleration alternatives

Service Name

Amazon CloudWatch

Service Overview

Amazon CloudWatch is a comprehensive monitoring and observability service that provides real-time visibility into AWS resources and applications. It collects and tracks metrics, monitors log files, and sets alarms to enable system-wide observability. CloudWatch offers operational visibility through metrics, alarms, and dashboards; application performance monitoring through Application Signals and Synthetics; infrastructure monitoring with specialized insights; comprehensive log management and analysis; and network monitoring capabilities. The service integrates seamlessly with AWS services and supports both automated and custom monitoring across multi-account environments.

Key Use Cases

- Infrastructure monitoring: Monitor AWS resources like EC2 instances, RDS databases, and EKS clusters with real-time metrics and performance data
- Application performance monitoring: Track application health, latency, error rates, and request volumes with automatic detection and visualization
- Log management and analysis: Centralize, store, search, and analyze logs from AWS services and custom applications using CloudWatch Logs Insights
- Operational alerting: Set up automated alarms based on thresholds to trigger notifications and automated responses to resource changes
- Cross-account observability: Monitor applications spanning multiple AWS accounts from a central monitoring account with unified dashboards and analysis

Features

- **Metrics and Alarms**: Collect, store, and track performance metrics with customizable alarms that trigger automated actions when thresholds are breached
- **Dashboards**: Create unified visualizations of metrics and logs with customizable dashboards that can be shared across accounts and regions
- **Application Signals**: Automatically detect and monitor application performance indicators like latency, error rates, and request rates without manual instrumentation
- **CloudWatch Logs**: Comprehensive log management with storage, search, analysis, and real-time processing capabilities using multiple query languages
- **Container Insights**: Specialized monitoring for containerized applications on Amazon ECS, EKS, and self-managed Kubernetes clusters

- **Database Insights**: Real-time database performance monitoring with SQL query analysis and load troubleshooting for AWS database services
- **Lambda Insights**: System-level metrics for Lambda functions including memory, CPU utilization, cold start detection, and performance analysis
- **Synthetics**: Proactive endpoint and API monitoring through configurable canaries that simulate user behavior and detect issues
- **Network Monitoring**: Internet Monitor and Network Flow Monitor for analyzing network performance, connectivity, and identifying performance bottlenecks
- **Cross-Account Observability**: Centralized monitoring across multiple AWS accounts with unified dashboards, alarms, and root-cause analysis capabilities

Value Proposition

Amazon CloudWatch provides native, deep integration with all AWS services, offering automatic metric collection and monitoring without additional setup. Unlike third-party solutions, CloudWatch eliminates the complexity of managing separate monitoring infrastructure while providing unified visibility across compute, storage, database, and network resources:

- The service offers comprehensive cost optimization through tiered pricing, free tiers for basic usage, and pay-as-you-consume models
- CloudWatch's cross-account observability, automated dashboards, and AI-powered insights reduce operational overhead while maintaining enterprise-scale reliability and security
- The seamless integration with AWS services enables automated responses and remediation actions directly within the AWS ecosystem

Service Integration

CloudWatch integrates natively with virtually all AWS services including Amazon EC2 for instance monitoring, Amazon RDS for database performance, AWS Lambda for function metrics, Amazon S3 for storage analytics, and Amazon EKS/ECS for container insights. It works closely with AWS Auto Scaling for automated resource adjustment, Amazon SNS for alarm notifications, AWS Systems Manager for centralized configuration, and AWS Organizations for multi-account monitoring:

- CloudWatch also integrates with AWS X-Ray for distributed tracing, AWS CloudTrail for API monitoring, Amazon EventBridge for event-driven automation, and AWS Backup for monitoring backup operations
- The service supports AWS IAM for access control and works with VPC endpoints for secure private connectivity

Onboarding and Offboarding

Getting started with CloudWatch requires an AWS account and basic IAM permissions. Many AWS services automatically send metrics to CloudWatch at no additional cost, providing immediate visibility:

- Users can access the CloudWatch console, view automatic dashboards for their resources, and set up custom alarms and dashboards as needed
- The CloudWatch agent can be installed on EC2 instances and on-premises servers for detailed system metrics
- For advanced features like Application Signals or Container Insights, users simply enable these services through the console
- CloudWatch provides getting started guides, solution catalog templates, and pre-built configurations to accelerate implementation across common use cases and AWS services

Getting Started Guide:
https://docs.aws.amazon.com/AmazonCloudWatch/latest/monitoring/GettingSetup.html

Data Removal

Data removal in CloudWatch involves configuring retention policies for logs, metrics, and alarms based on business requirements. CloudWatch automatically expires metrics after 15 months, but users can delete custom metrics, alarms, and dashboards immediately. Log data can be deleted by adjusting log group retention policies or deleting entire log groups. When offboarding, customers should disable CloudWatch agents, remove custom metrics publishing, delete dashboards and alarms, and clear log groups to stop incurring charges while ensuring compliance with data retention requirements.

Backup/Restore and Disaster Recovery

CloudWatch itself provides monitoring for backup and disaster recovery operations rather than direct backup capabilities. The service monitors AWS Backup operations, tracks backup success/failure metrics, and can alert on backup issues. CloudWatch supports cross-region monitoring for disaster recovery scenarios and integrates with AWS Elastic Disaster Recovery for monitoring replication health. Metric and log data are automatically replicated across multiple Availability Zones for high availability, and cross-region log centralization enables backup of log data to alternate regions for compliance and disaster recovery purposes.

Backup Capabilities

CloudWatch does not provide traditional backup capabilities for applications or data. Instead, it serves as the monitoring layer for backup operations performed by other AWS services like AWS Backup, Amazon EBS snapshots, and RDS automated backups:

- CloudWatch can monitor backup job success rates, alert on backup failures, and track backup storage utilization
- The service stores its own metrics and logs with built-in durability and retention policies, but users should configure appropriate retention settings and cross-region log centralization for long-term data preservation and compliance requirements

Disaster Recovery

CloudWatch supports disaster recovery scenarios through cross-region monitoring capabilities, centralized logging, and integration with AWS disaster recovery services. The service can monitor the health of disaster recovery infrastructure, track replication metrics, and provide visibility into failover operations:

- CloudWatch works with AWS Elastic Disaster Recovery by monitoring replication lag and recovery point objectives
- Cross-account observability enables monitoring of disaster recovery environments in separate accounts, while automated alarms can trigger disaster recovery procedures when primary systems fail or performance degrades below acceptable thresholds

Recovery Objectives

CloudWatch helps customers meet RPO and RTO objectives by providing real-time monitoring of backup operations, replication health, and system availability. The service can monitor backup frequency and success rates to ensure RPO compliance, while tracking application response times and availability metrics supports RTO requirements. CloudWatch alarms can automatically trigger disaster recovery procedures when systems fail, reducing recovery time. Cross-region monitoring enables tracking of disaster recovery infrastructure health, and integration with AWS Elastic Disaster Recovery provides visibility into replication lag and recovery readiness to maintain desired recovery objectives.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/AmazonCloudWatch/latest/monitoring/cloudwatch_limits.html

FAQ: https://aws.amazon.com/cloudwatch/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/AmazonCloudWatch/latest/monitoring/WhatIsCloudWatch.html

User Guide: https://docs.aws.amazon.com/AmazonCloudWatch/latest/monitoring/

API Reference: https://docs.aws.amazon.com/AmazonCloudWatch/latest/APIReference/

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/cloudwatch/pricing/

Cost Optimization

CloudWatch offers several unique cost optimization features including tiered pricing for vended logs starting at $0.50/GB and decreasing to $0.05/GB at higher volumes, free tier allowances for basic metrics and log ingestion, and volume discounts for custom metrics. The service provides cost analysis tools to identify high-cost log groups and metrics, supports log retention policies to automatically delete old data, and offers multiple log destinations (S3, Kinesis Data Firehose) at lower costs than CloudWatch Logs storage. Customers can optimize costs by using metric filters instead of custom metrics, implementing log sampling, rightsizing alarm frequencies, and leveraging cross-account observability to consolidate monitoring infrastructure across multiple accounts.

Savings Considerations

Primary cost savings opportunities include configuring appropriate log retention policies to avoid indefinite storage costs, using log sampling and filtering to reduce ingestion volumes, and leveraging the free tier for basic monitoring needs. Customers should optimize custom metrics by reducing cardinality and publication frequency, use CloudWatch agent filtering to send only necessary metrics, and consider alternative storage options like S3 for long-term log retention:

- Consolidating monitoring through cross-account observability reduces duplicate infrastructure costs, while using pre-built dashboards and solution catalog templates minimizes development overhead
- Regular cost analysis using AWS Cost Explorer helps identify and eliminate unused alarms, dashboards, and log groups that continue accruing charges

Service Name

Amazon CodeGuru

Service Overview

Amazon CodeGuru is a developer tool that provides intelligent recommendations to improve code quality and identify an application's most expensive lines of code. It consists of two components: Amazon CodeGuru Reviewer and Amazon CodeGuru Profiler. CodeGuru Reviewer uses machine learning and automated reasoning to analyze code during development, detecting critical issues, security vulnerabilities, and bugs in Java and Python code. CodeGuru Profiler continuously analyzes application runtime performance, identifying expensive lines of code and performance bottlenecks in production environments to reduce operational costs and improve efficiency.

Key Use Cases

- Automated code review and quality improvement: Detect critical defects, security vulnerabilities, and coding best practice deviations in pull requests
- Performance optimization: Identify expensive lines of code and CPU bottlenecks in production applications to reduce infrastructure costs
- Security vulnerability detection: Scan code for OWASP Top 10 vulnerabilities, AWS API security best practices, and sensitive information leaks

Features

- **CodeGuru Reviewer**: Automated code review service that uses machine learning to detect critical issues, security vulnerabilities, and deviations from best practices in Java and Python code
- **CodeGuru Profiler**: Performance profiling tool that continuously analyzes application behavior in production with minimal overhead to identify expensive lines of code
- **Interactive Flame Graphs**: Visual representations of code paths that consume the most resources, helping understand performance bottlenecks
- **Anomaly Detection**: Real-time analysis of application profiles to detect anomalies in behavior and methods with SNS notifications
- **Security Detector**: Specialized component that focuses on detecting security vulnerabilities using static analysis and control-flow graphs
- **CI/CD Integration**: Integration with GitHub Actions and other CI/CD providers to provide recommendations within pull requests

Value Proposition

Amazon CodeGuru offers unique machine learning-powered code analysis trained on millions of code reviews from major open-source projects and Amazon's codebase. Unlike traditional static analysis tools, it provides intelligent, context-aware recommendations with low false positive rates:

- The service combines automated code review during development with continuous production profiling, offering a comprehensive solution for both code quality and performance optimization
- Key advantages include minimal performance impact during profiling, seamless integration with existing development workflows, and the ability to quantify the cost impact of inefficient code, enabling data-driven optimization decisions

Service Integration

CodeGuru integrates deeply with core AWS developer services and infrastructure. For code repositories, it supports AWS CodeCommit, GitHub, GitHub Enterprise, and Bitbucket through Amazon CodeConnections:

- It integrates with AWS Lambda, Amazon EC2, Amazon ECS, Amazon EKS, AWS Fargate, and AWS Elastic Beanstalk for application profiling
- The service works with AWS CodeBuild, AWS CodePipeline for CI/CD workflows, and AWS Secrets Manager for unprotected secrets detection
- CloudWatch integration provides monitoring and alerting capabilities, while IAM manages access controls
- Amazon SNS delivers anomaly notifications, and AWS CloudTrail provides audit logging for all CodeGuru activities

Onboarding and Offboarding

Customers get started with CodeGuru through the AWS Management Console by creating profiling groups for CodeGuru Profiler or associating repositories for CodeGuru Reviewer. For Profiler, users configure a profiling group, start their application with the profiling agent, and begin receiving performance recommendations:

- For Reviewer, customers associate their code repositories (AWS CodeCommit, GitHub, Bitbucket) with the service, after which it automatically analyzes pull requests and provides recommendations
- The setup process is streamlined with one-time configuration, and the service includes comprehensive documentation, tutorials, and sample applications to accelerate adoption across development teams

Getting Started Guide: https://docs.aws.amazon.com/codeguru/latest/profiler-ug/getting-started.html

Data Removal

CodeGuru does not persistently store source code, accessing it only during analysis. For CodeGuru Profiler, customers can stop the profiling agent and delete profiling groups to cease data collection. Repository associations can be disassociated to stop CodeGuru Reviewer analysis. Recommendation reports and profiling data are retained for 30 days. Customers should remove IAM service-linked roles and policies, and clean up any CloudWatch metrics or SNS notifications configured for the service.

Backup/Restore and Disaster Recovery

Amazon CodeGuru is a managed AWS service where AWS handles backup and disaster recovery of the underlying infrastructure and service components. The service operates across multiple AWS regions with built-in redundancy. Recommendation reports and profiling data are automatically replicated and maintained by AWS. Customers do not need to implement separate backup strategies for CodeGuru service data, as AWS ensures service availability and data durability through the managed service architecture.

Backup Capabilities

CodeGuru does not have traditional backup capabilities as it is a managed analysis service. The service does not require AWS Backup integration since it doesn't store persistent customer application data:

- Profiling data and recommendations are maintained by AWS with automatic redundancy
- Customers should focus on backing up their source code repositories and application deployments using appropriate services for their code hosting platforms

Disaster Recovery

CodeGuru is a multi-region managed service that does not require AWS Elastic Disaster Recovery. AWS manages the disaster recovery for the service infrastructure automatically:

- The service is available in multiple regions, allowing customers to use CodeGuru in their preferred regions for analysis
- Service continuity is maintained through AWS's managed service architecture, with automatic failover and redundancy handled transparently by AWS

Recovery Objectives

CodeGuru supports customer recovery objectives indirectly by helping build more resilient applications through code quality improvements and performance optimization. By detecting code defects, security vulnerabilities, and performance issues before deployment, the service reduces the likelihood of production failures that would impact RPO and RTO. The continuous profiling capabilities help maintain optimal application

performance, contributing to better system reliability and faster recovery from performance-related issues.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/codeguru/latest/reviewer-ug/quotas.html

FAQ: https://aws.amazon.com/codeguru/profiler/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/codeguru/latest/reviewer-ug/welcome.html

User Guide: https://docs.aws.amazon.com/codeguru/latest/profiler-ug/what-is-codeguru-profiler.html

API Reference: https://docs.aws.amazon.com/codeguru/latest/reviewer-api/API_ListCodeReviews.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/codeguru/pricing/

Cost Optimization

CodeGuru offers unique cost optimization through intelligent identification of expensive code lines and performance bottlenecks. The Profiler component estimates the actual cost impact of inefficient code, enabling data-driven optimization decisions:

- Customers can reduce infrastructure costs by optimizing CPU utilization and eliminating performance bottlenecks identified by the service
- The 90-day free tier for both components (100K lines of code for Reviewer, 36,000 sampling hours for Profiler) allows cost-effective evaluation
- The service's ability to prevent costly production issues through early detection of code defects provides additional cost savings

Savings Considerations

Primary cost savings come from reduced infrastructure requirements through performance optimization and decreased operational overhead from fewer production issues. CodeGuru helps teams identify and eliminate expensive code patterns before they impact production costs:

- The service reduces the need for manual code reviews and performance analysis, saving developer time and resources
- By detecting security vulnerabilities early, it prevents potential costly security incidents
- The continuous profiling capability helps maintain optimal resource utilization, preventing over-provisioning and reducing cloud infrastructure costs over time
- Organizations typically see ROI through reduced compute costs and improved developer productivity

Service Name

Amazon Cognito

Service Overview

Amazon Cognito is a customer identity and access management (CIAM) service that provides user authentication, authorization, and user management for web and mobile applications. It consists of two main components: user pools for authentication and identity management, and identity pools for authorization and AWS resource access. The service enables developers to add user sign-up, sign-in, access control, and AWS service access capabilities to applications with secure, scalable identity stores that can handle millions of users. Amazon Cognito supports various authentication methods including social identity providers, SAML/OIDC federation, multi-factor authentication, and passwordless authentication options like passkeys.

Key Use Cases

- User authentication and authorization: Implementing secure sign-up, sign-in, and access control for web and mobile applications with customizable UI and authentication flows
- Social and enterprise identity federation: Enabling users to authenticate through social providers (Google, Facebook, Amazon, Apple) or enterprise identity providers using SAML 2.0 and OIDC standards
- AWS resource access control: Granting temporary AWS credentials to authenticated users through identity pools for fine-grained access to AWS services like S3, DynamoDB, and Lambda

Features

- **User Pools**: User directories that handle sign-up, sign-in, and user profile management with customizable authentication workflows
- **Identity Pools**: Authorization service that grants temporary AWS credentials to authenticated or anonymous users for accessing AWS resources
- **Managed Login**: No-code, branded web-based sign-in interface with customizable UI and support for multiple authentication methods
- **Multi-Factor Authentication**: Support for SMS, email OTP, authenticator apps, and WebAuthn passkeys for enhanced security
- **Federation Support**: Integration with external identity providers including social providers and SAML/OIDC-based enterprise systems
- **Lambda Triggers**: Customizable user workflows and authentication flows using serverless functions for extensibility

- **Machine-to-Machine Authentication**: OAuth 2.0 client credentials flow for API-to-API authentication with token customization
- **Risk-Based Authentication**: Adaptive authentication based on device, location, and behavioral analysis with compromised credential protection

Value Proposition

Amazon Cognito provides a comprehensive, fully managed identity solution that eliminates the complexity of building authentication systems from scratch. It offers enterprise-grade security with compliance certifications (HIPAA, PCI DSS, SOC, ISO), built-in protection against common threats, and advanced security features like adaptive authentication:

- The service scales seamlessly to millions of users with high availability across multiple AWS regions
- Cost-effective pricing based on monthly active users with generous free tiers makes it accessible for startups while supporting enterprise needs
- Deep integration with AWS services and extensive SDK support accelerates development, while features like managed login and Lambda triggers provide flexibility without sacrificing security

Service Integration

Amazon Cognito integrates natively with core AWS services to provide comprehensive application security and functionality. API Gateway uses Cognito user pools as authorizers for REST APIs, enabling seamless authentication and authorization:

- Lambda functions can be triggered by Cognito events for custom authentication flows and user management workflows
- Identity pools grant temporary AWS credentials for direct access to services like S3, DynamoDB, and SNS
- Integration with AWS WAF protects managed login pages from web vulnerabilities
- CloudTrail provides audit logging, while CloudWatch offers monitoring and analytics
- AWS Amplify provides simplified integration for frontend frameworks, and AWS Verified Permissions enables fine-grained authorization policies
- Additionally, Cognito works with AWS PrivateLink for private connectivity and integrates with AWS Secrets Manager for secure credential storage

Onboarding and Offboarding

Customers can get started with Amazon Cognito through multiple guided setup options in the AWS Console. The new streamlined console experience offers a quick setup wizard with use case-specific recommendations and framework-specific integration instructions:

- Developers can choose from three main approaches: using managed login for no-code implementation, integrating SDKs for custom authentication interfaces, or leveraging AWS Amplify for full-stack applications
- The service provides extensive code examples, workshops through AWS Workshop Studio, and starter templates for popular frameworks
- Getting started involves creating a user pool, configuring authentication methods, setting up an app client, and integrating with applications using provided SDKs and documentation

Getting Started Guide:
https://docs.aws.amazon.com/cognito/latest/developerguide/cognito-guided-setup.html

Data Removal

Customers can remove user data through the DeleteUser API operation or console interface, which permanently deletes user accounts and associated profile data from user pools. User pools and identity pools can be deleted entirely through the console or APIs, removing all contained user data. For compliance requirements, customers can export user data before deletion. The service provides granular control over data retention and supports bulk user deletion for offboarding scenarios.

Backup/Restore and Disaster Recovery

Amazon Cognito provides high availability and automatic data replication across multiple Availability Zones within regions but does not offer traditional backup and restore capabilities. The service maintains user data durability through AWS's infrastructure redundancy. For disaster recovery, customers should implement user pool configuration backups and consider multi-region deployments for critical applications.

Backup Capabilities

Amazon Cognito does not integrate with AWS Backup service and does not provide built-in backup capabilities for user data or configurations. The service relies on AWS infrastructure redundancy for data protection. Customers must implement their own backup strategies for user pool configurations and consider exporting user data if needed.

Disaster Recovery

Amazon Cognito does not integrate with AWS Elastic Disaster Recovery service. The service provides regional high availability but customers must implement cross-region disaster recovery strategies manually. This includes replicating user pool configurations and potentially migrating user data between regions during disaster scenarios.

Recovery Objectives

Amazon Cognito helps customers meet RPO and RTO objectives through its multi-AZ architecture within regions, providing automatic failover and high availability. However, for

cross-region recovery, customers must implement manual processes including user pool recreation and data migration, which may result in longer RTOs depending on implementation complexity and data volumes.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/cognito/latest/developerguide/quotas.html

FAQ: https://aws.amazon.com/cognito/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/cognito/latest/developerguide/what-is-amazon-cognito.html

User Guide: https://docs.aws.amazon.com/cognito/latest/developerguide/what-is-amazon-cognito.html

API Reference: https://docs.aws.amazon.com/cognito-user-identity-pools/latest/APIReference/

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/cognito/pricing/

Cost Optimization

Amazon Cognito offers several unique cost optimization opportunities through its Monthly Active User (MAU) pricing model. The service provides generous free tiers with 10,000 MAUs for Essentials and Lite tiers, and 50 MAUs for federated authentication:

- Cost optimization strategies include avoiding activation of inactive users, linking federated users to local profiles to reduce federation costs, managing request rates to stay within default quotas, implementing proper token refresh mechanisms to reduce unnecessary API calls, and using ListUsers instead of AdminGetUser for bulk operations
- The recent removal of machine-to-machine app client pricing makes M2M authentication more cost-effective
- Customers can separate users with different feature requirements into different user pools to optimize tier selection

Savings Considerations

Cost savings can be achieved through strategic user pool architecture and feature tier selection. The Lite tier offers significant savings for basic password authentication use cases, while the Plus tier provides up to 60% savings compared to standalone advanced security features:

- Implementing efficient token management reduces API call costs, and proper user lifecycle management prevents billing for inactive users
- Federated authentication can be cost-optimized by linking external identities to local profiles
- The simplified M2M pricing model eliminates app client charges, reducing costs for API authentication scenarios
- Using AWS Free Tier benefits and monitoring usage through the Service Quotas console helps optimize spending
- Customers should regularly review their feature usage to ensure they're on the most cost-effective tier for their needs

Service Name

Amazon Connect

Service Overview

Amazon Connect is a cloud-based omnichannel contact center solution that enables businesses to deliver high-quality customer experiences across voice, chat, SMS, email, web calling, video, and tasks. It provides AI-powered capabilities including conversational IVR, chatbots, agent assistance, real-time analytics, and workforce management. Amazon Connect offers a pay-as-you-go pricing model with no upfront costs, minimum fees, or long-term commitments, allowing businesses to scale their contact center operations up or down based on demand while maintaining seamless customer experiences across all channels.

Key Use Cases

- Customer service and support: Handle inbound customer inquiries, complaints, and support requests across voice, chat, email, and SMS channels with intelligent routing and AI assistance
- Sales and lead generation: Manage outbound campaigns with predictive dialing, answering machine detection, and ML-powered optimization for appointment reminders, marketing promotions, and follow-ups
- Self-service automation: Deploy AI-powered chatbots and conversational IVR using Amazon Lex for customer authentication, order status, FAQ responses, and routine transactions without agent intervention
- Technical support and IT help desk: Provide multi-tier technical support with case management, skill-based routing, and knowledge base integration for complex problem resolution
- Collections and payment processing: Conduct secure payment collection, account management, and debt recovery operations with compliance features and automated workflows
- Emergency services and crisis management: Handle high-volume emergency calls with priority routing, overflow management, and real-time dashboards for critical response coordination

Features

- **Omnichannel Communication**: Voice, chat, email, SMS, video calling, and task management in a unified platform
- **AI-Powered Self-Service**: Conversational IVR and chatbots with Amazon Lex for natural language interactions

- **Agent Workspace**: Integrated agent desktop with unified customer profiles, step-by-step guides, and real-time assistance
- **Contact Routing**: Skills-based routing, queue management, and intelligent contact distribution based on agent capabilities
- **Real-time Analytics**: Live dashboards, historical reports, conversational analytics, and sentiment analysis
- **Workforce Management**: ML-powered forecasting, capacity planning, agent scheduling, and adherence tracking
- **Outbound Campaigns**: High-volume predictive dialing with answering machine detection and campaign optimization
- **Generative AI Assistant**: Amazon Q in Connect provides real-time agent recommendations and post-contact summaries
- **Drag-and-Drop Flow Designer**: Visual workflow designer for creating personalized customer experiences across all channels
- **Voice Quality & Telephony**: High-quality voice calls, global telephony coverage, and telephony-as-a-service model

Value Proposition

Amazon Connect delivers superior value through its cloud-native architecture that eliminates infrastructure management while providing unlimited AI capabilities at predictable pricing. Unlike traditional contact center solutions, it offers seamless omnichannel experiences with automatic scaling, pay-as-you-go pricing, and no long-term commitments. Key differentiators include: integrated AWS ecosystem allowing seamless integration with 200+ AWS services, built-in generative AI that enhances both agent productivity and customer self-service without additional complexity, true omnichannel design that maintains context across voice, chat, email, and tasks, enterprise-grade security and compliance built on AWS infrastructure, and rapid deployment with pre-built integrations and drag-and-drop configuration tools that reduce implementation time from months to weeks.

Service Integration

Amazon Connect integrates deeply with the AWS ecosystem for comprehensive contact center functionality. Core integrations include Amazon Lex for conversational AI and chatbots, Amazon Polly for text-to-speech, Amazon Transcribe for call transcription, Amazon Comprehend for sentiment analysis, AWS Lambda for custom business logic and external system integrations, Amazon S3 for call recording and report storage, Amazon DynamoDB for real-time data access and contact routing, Amazon Kinesis for real-time data streaming and analytics, Amazon CloudWatch for monitoring and alerting, AWS Identity and Access Management (IAM) for security and permissions, Amazon Pinpoint for omnichannel messaging, Amazon SNS for notifications, and AWS Directory Service for user federation. Additionally, it leverages Amazon EventBridge for event-driven workflows and AWS CloudFormation for infrastructure as code deployment.

Onboarding and Offboarding

Getting started with Amazon Connect follows a streamlined six-step process. First, create an Amazon Connect instance in the AWS Management Console, specifying user management preferences, telephony options, and data storage location:

- Second, claim phone numbers from AWS inventory or port existing numbers for voice capabilities
- Third, set up routing by creating queues, routing profiles, and defining hours of operation for different channels
- Fourth, design customer flows using the drag-and-drop flow designer to define the end-to-end customer experience
- Fifth, add users including managers and agents, configuring their routing profiles, security permissions, and equipment settings
- Finally, configure chat tools and widgets for web integration if using digital channels
- The entire setup can be completed in hours, with AWS providing tutorials, workshops, and guided setup wizards to accelerate deployment

Getting Started Guide:
https://docs.aws.amazon.com/connect/latest/adminguide/amazon-connect-get-started.html

Data Removal

Offboarding from Amazon Connect involves deleting the instance through the AWS Management Console, which releases phone numbers back to inventory and makes the service unavailable to customers. Upon deletion, all instance settings, data, metrics, and reports become permanently inaccessible and cannot be restored. Scheduling data is retained for 30 days for GDPR compliance before final deletion. Customers must manually delete associated CloudWatch log groups if flow logging was enabled, and any stored call recordings in S3 buckets must be managed separately according to retention policies.

Backup/Restore and Disaster Recovery

Amazon Connect provides built-in resilience and data protection through AWS's global infrastructure with active-active architecture across Availability Zones. Contact records, flow configurations, and user settings are automatically replicated across multiple AZs for high availability. Call recordings and reports stored in S3 benefit from 99.999999999% durability and can leverage S3 Cross-Region Replication for additional protection. Flow configurations can be exported and version-controlled for backup and restoration purposes. However, Amazon Connect does not currently integrate with AWS Backup service for centralized backup management.

Backup Capabilities

Amazon Connect does not directly support AWS Backup integration for automated backup management. However, the service provides alternative data protection mechanisms including automatic replication of configurations and data across Availability Zones, S3-based storage for call recordings with built-in durability and optional cross-region replication, exportable flow configurations for version control and backup purposes, and API-based configuration backup through programmatic export of settings, users, queues, and routing profiles. Contact records and analytics data can be streamed to external systems for backup and archival purposes.

Disaster Recovery

Amazon Connect does not integrate with AWS Elastic Disaster Recovery service but provides inherent disaster recovery capabilities through its cloud-native, multi-AZ architecture. The service automatically handles failover between Availability Zones with no customer intervention required:

- For comprehensive disaster recovery, customers can implement cross-region strategies by setting up instances in multiple regions, exporting and replicating flow configurations, implementing DNS-based failover with Route 53, and leveraging S3 Cross-Region Replication for call recordings
- The service's API-driven nature enables Infrastructure as Code approaches for rapid environment recreation

Recovery Objectives

Amazon Connect helps customers achieve aggressive RPO and RTO objectives through its cloud-native architecture. The service provides near-zero RPO for real-time operations through automatic multi-AZ replication and continuous data synchronization. RTO is minimized through automatic failover capabilities within the same region, typically achieving recovery in minutes. For cross-region scenarios, customers can design solutions with RTOs of 15-30 minutes by pre-deploying instances and using automation for configuration replication. The service's elastic scaling ensures capacity is available during disaster recovery scenarios without pre-provisioning requirements.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/connect/latest/adminguide/amazon-connect-service-limits.html

FAQ: https://aws.amazon.com/connect/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/connect/latest/adminguide/

User Guide: https://docs.aws.amazon.com/connect/latest/adminguide/

API Reference: https://docs.aws.amazon.com/connect/latest/APIReference/

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/connect/pricing/

Cost Optimization

Amazon Connect offers several unique cost optimization strategies beyond standard AWS practices. Region selection significantly impacts telephony costs for claimed phone numbers and inbound usage, with potential savings of 20-40% by choosing optimal regions:

- Callback functionality reduces costs by allowing customers to opt for callbacks during high-volume periods instead of waiting on hold, decreasing handle times and PSTN charges
- Self-service implementation using Amazon Lex reduces agent labor costs by handling routine inquiries automatically
- Redirecting voice contacts to chat allows agents to handle multiple conversations simultaneously, improving efficiency
- Using click-to-call reduces call durations by passing contextual information
- The unlimited AI pricing model provides predictable costs for all optimization features, while softphone usage eliminates deskphone PSTN extension costs

Savings Considerations

Primary cost savings with Amazon Connect stem from its pay-as-you-go model eliminating upfront license fees, infrastructure costs, and maintenance expenses typically associated with traditional contact centers. The service's elastic scaling means customers only pay for actual usage rather than peak capacity, potentially saving 30-60% compared to traditional solutions:

- Studies show organizations achieve $4.3 million in cloud technology cost savings, 31% subscription cost savings, and $4.6 million in agent labor savings through improved productivity
- Additional savings come from reduced IT overhead, elimination of hardware refresh cycles, automatic software updates, and improved agent efficiency through AI-powered tools
- Storage cost optimization through S3 lifecycle policies can transition old recordings to cheaper storage tiers, while strategic use of self-service capabilities reduces operational costs per contact

Service Name

Amazon DataZone

Service Overview

Amazon DataZone is a data management service that enables organizations to catalog, discover, share, and govern data across AWS, on-premises, and third-party sources. It provides fine-grained access controls to ensure secure data access and compliance with organizational regulations while simplifying data architecture through integration with various AWS services including Amazon Redshift, Amazon Athena, AWS Glue, and AWS Lake Formation. The service features a business data catalog, governed data sharing workflows, and a browser-based data portal for collaboration across teams.

Key Use Cases

- Breaking down data silos: Enable business teams to build domains and business data catalogs while maintaining control over access and making data more discoverable across organizational boundaries.
- Data mesh architecture implementation: Support decentralized data ownership and federated governance model where data producers securely share data and context while consumers can access needed data through approval workflows.
- Automated data discovery and cataloging: Reduce manual data entry and improve searching experiences by automating metadata generation and enrichment using AI-powered recommendations for better data understanding.

Features

- **Business Data Catalog**: Centralized catalog with AI-generated metadata, business context descriptions, and automated recommendations for data assets
- **Governed Data Sharing**: Subscription and fulfillment workflows with automated permission management for AWS Glue and Amazon Redshift resources
- **Projects and Environments**: Collaborative workspaces with role-based access controls and integration with AWS services through blueprints
- **Data Products**: Collections of interconnected data assets designed for specific business use-cases with simplified administration
- **Fine-grained Access Control**: Row and column-level filtering using AWS Lake Formation Data Cell Filters and Amazon Redshift late binding views
- **Multi-account Integration**: Support for data sharing across multiple AWS accounts with centralized governance and domain organization

Value Proposition

Amazon DataZone provides a unified solution for data governance that combines cataloging, discovery, sharing, and governance in a single service, reducing complexity compared to managing multiple separate tools. It offers automated AI-powered metadata enrichment, seamless integration with existing AWS analytics services, and supports both centralized and decentralized data mesh architectures. The service enables self-service data access while maintaining security and compliance, reducing time-to-value for data analytics initiatives and eliminating manual data management overhead.

Service Integration

Amazon DataZone integrates natively with AWS Glue Data Catalog and Amazon Redshift as producer data sources for automatic metadata ingestion. It connects with Amazon Athena and Amazon Redshift Query Editor for direct data analysis, Amazon SageMaker for machine learning workflows, and AWS Lake Formation for permission management. The service also integrates with Amazon QuickSight for visualization, AWS IAM Identity Center for authentication, and supports custom integrations through standard events for other AWS services and third-party solutions.

Onboarding and Offboarding

Customers start by configuring AWS IAM Identity Center and Lake Formation permissions, then create an Amazon DataZone domain through the management console. The getting started process includes creating publishing projects, environment profiles, and environments using built-in blueprints for data lakes, data warehouses, or Amazon SageMaker. Sample scripts and quickstart workflows are available for AWS Glue and Amazon Redshift data sources to automate initial setup and demonstrate core workflows for data producers and consumers.

Getting Started Guide: https://docs.aws.amazon.com/datazone/latest/userguide/getting-started.html

Data Removal

Data removal in Amazon DataZone involves deleting assets from project inventories, unpublishing data from the catalog, removing subscription grants, and deleting projects and environments. Domain administrators can delete the entire domain to remove all associated data. The service follows AWS data retention policies, and customers can implement automated data lifecycle policies for underlying storage services like Amazon S3 to manage long-term data retention and deletion.

Backup/Restore and Disaster Recovery

Amazon DataZone metadata is stored in AWS managed infrastructure with built-in redundancy and durability. The service does not provide explicit backup features as it

primarily manages metadata and access permissions rather than the underlying data. Data protection relies on the backup capabilities of the underlying data stores like Amazon S3, AWS Glue Data Catalog, and Amazon Redshift that Amazon DataZone integrates with.

Backup Capabilities

Amazon DataZone does not have dedicated backup capabilities as it manages metadata and access controls rather than storing the actual data. The service relies on AWS managed infrastructure for metadata durability. Customers should implement backup strategies for the underlying data sources that Amazon DataZone catalogs, such as using AWS Backup for supported services or native backup features of Amazon Redshift and Amazon RDS.

Disaster Recovery

Amazon DataZone is a regional service that relies on AWS infrastructure resilience for disaster recovery. The service does not integrate directly with AWS Elastic Disaster Recovery as it manages metadata rather than primary data workloads. Disaster recovery planning should focus on the underlying data sources and analytics infrastructure that Amazon DataZone manages, using appropriate AWS disaster recovery services for those components.

Recovery Objectives

Amazon DataZone helps with recovery objectives by maintaining metadata consistency and access control definitions that can be quickly restored when underlying data sources recover. The service's metadata catalog provides essential information for data recovery processes, including data lineage, business context, and access permissions. Recovery time is primarily dependent on the underlying data services rather than Amazon DataZone itself.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/datazone/latest/userguide/datazone-limits.html

FAQ: https://aws.amazon.com/datazone/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/datazone/latest/userguide/what-is-datazone.html

User Guide: https://docs.aws.amazon.com/datazone/latest/userguide/getting-started.html

API Reference: https://docs.aws.amazon.com/datazone/latest/APIReference/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/datazone/pricing/

Cost Optimization

Amazon DataZone offers several cost optimization features including a pay-as-you-go pricing model with no upfront costs or user-level subscription fees as of November 2024. The service provides free usage tiers for requests (4,000 per month), metadata storage (20 MB), and compute (0.2 units), plus free access to core APIs for domain and project management. Cost optimization comes from reducing data duplication by enabling data discovery and sharing, eliminating manual data management overhead, and providing automated metadata generation to reduce operational costs.

Savings Considerations

Primary cost savings come from eliminating data silos and reducing duplicate data storage across teams through improved data discovery and sharing. The service reduces operational overhead by automating metadata management and data cataloging processes that would otherwise require manual effort:

- Organizations can achieve cost savings by implementing data mesh architectures that eliminate the need for centralized data lake infrastructure while maintaining governance
- The removal of per-user subscription fees makes the service more cost-effective for large-scale deployments with many users

Service Name

Amazon Data Firehose

Service Overview

Amazon Data Firehose is a fully managed service for delivering real-time streaming data to destinations such as Amazon S3, Amazon Redshift, Amazon OpenSearch Service, Splunk, Apache Iceberg Tables, and custom HTTP endpoints. The service automatically buffers, transforms, compresses, and delivers streaming data without requiring customers to write applications or manage resources. It supports data ingestion from multiple sources including Amazon Kinesis Data Streams, Amazon MSK, Direct PUT API, and AWS services like CloudWatch Logs. Data Firehose automatically scales to match the data throughput and provides near real-time data delivery with configurable buffer sizes and intervals.

Key Use Cases

- Streaming data into data lakes and warehouses: Loading real-time data from applications into Amazon S3 for analytics and long-term storage
- Security monitoring and threat detection: Streaming security logs and events to security analytics platforms like Splunk for real-time monitoring
- Machine learning data pipelines: Feeding streaming data into ML platforms and analytics services for real-time inference and model training
- Business intelligence and analytics: Delivering streaming business metrics to data warehouses like Amazon Redshift for real-time dashboards
- Log aggregation and analysis: Centralizing application and infrastructure logs for monitoring, troubleshooting, and compliance

Features

- **Fully Managed Service**: No infrastructure management required, automatic scaling and provisioning
- **Multiple Data Sources**: Supports Kinesis Data Streams, Amazon MSK, Direct PUT, and AWS service logs
- **Data Transformation**: Real-time data transformation using AWS Lambda functions before delivery
- **Format Conversion**: Convert data to columnar formats like Apache Parquet and ORC for optimized analytics
- **Dynamic Partitioning**: Automatically partition data in S3 based on content for improved query performance
- **Multiple Destinations**: Deliver to Amazon S3, Redshift, OpenSearch, Splunk, Snowflake, and HTTP endpoints

- **Buffer Management**: Configurable buffer sizes and intervals for optimized delivery batches
- **Encryption and Security**: Server-side encryption with AWS KMS and IAM-based access control
- **Error Handling**: Automatic retry logic and error logging with failed record backup
- **Monitoring Integration**: CloudWatch metrics and logging for stream monitoring and troubleshooting

Value Proposition

Amazon Data Firehose eliminates the complexity of building and managing streaming data pipelines by providing a fully managed, serverless solution. Customers choose Firehose for its zero infrastructure management, automatic scaling, and pay-as-you-go pricing model with no upfront commitments:

- The service reduces time-to-value by enabling immediate data streaming without writing custom applications or provisioning resources
- It offers built-in reliability features including automatic retries, error handling, and data backup capabilities
- Integration with over 30 AWS services and third-party destinations provides flexibility for diverse analytics architectures
- The service supports real-time data transformations and format conversions, optimizing data for downstream analytics while maintaining high availability and durability

Service Integration

Amazon Data Firehose integrates deeply with core AWS services across the data analytics ecosystem. It connects with Amazon Kinesis Data Streams and Amazon MSK as streaming data sources, while supporting direct ingestion from AWS services like CloudWatch Logs, VPC Flow Logs, and AWS WAF:

- For data processing, it integrates with AWS Lambda for real-time transformations and AWS Glue for metadata management and schema discovery
- Destination integrations include Amazon S3 for data lakes, Amazon Redshift for data warehousing, and Amazon OpenSearch Service for search analytics
- Security integrations include AWS KMS for encryption and AWS IAM for access control
- Monitoring is provided through Amazon CloudWatch for metrics and logging, while AWS Secrets Manager handles credential management for external destinations

Onboarding and Offboarding

Customers get started with Amazon Data Firehose by creating a delivery stream through the AWS Management Console, CLI, or SDKs. The process involves selecting a data source (Direct PUT, Kinesis Data Streams, or Amazon MSK), configuring optional data transformations, choosing a destination, and setting buffer configurations:

- Prerequisites include signing up for an AWS account and creating appropriate IAM roles for service permissions
- The service provides tutorials and sample code for common integration patterns
- Getting started is streamlined through the console wizard that guides users through stream creation, transformation setup, and destination configuration
- No infrastructure provisioning or application development is required, allowing customers to begin streaming data within minutes of setup

Getting Started Guide: https://docs.aws.amazon.com/firehose/latest/dev/what-is-this-service.html

Data Removal

Offboarding from Amazon Data Firehose involves stopping data ingestion, allowing existing buffered data to be delivered, and then deleting the delivery stream. Customers can pause data delivery temporarily or permanently delete streams through the console or APIs. Data removal depends on destination configurations - for S3 destinations, customers must manually delete objects from buckets; for other destinations like Redshift or OpenSearch, data persists according to those services' retention policies. Source data backup in S3 requires separate cleanup. Failed records stored in error buckets also need manual deletion during offboarding.

Backup/Restore and Disaster Recovery

Amazon Data Firehose provides built-in backup capabilities by optionally storing source data in Amazon S3 concurrently with destination delivery. This includes failed record backup for troubleshooting and source record backup for data transformation scenarios. The service offers at-least-once delivery semantics with automatic retry logic for up to 24 hours for DirectPut sources. For disaster recovery, Firehose relies on underlying AWS service durability and cross-region replication capabilities of destinations like S3. The service itself is highly available within regions but does not provide native cross-region replication.

Backup Capabilities

Amazon Data Firehose includes automatic backup features for data reliability and troubleshooting. Source data backup is available when using data transformations, storing original records in a separate S3 bucket:

- Failed record backup automatically saves records that couldn't be delivered to destinations in designated S3 error folders
- The service buffers data for up to 24 hours during delivery failures for DirectPut sources
- However, Firehose does not integrate directly with AWS Backup service as it's primarily a data streaming rather than storage service

Disaster Recovery

Amazon Data Firehose does not directly integrate with AWS Elastic Disaster Recovery as it's a data streaming service rather than compute infrastructure. The service provides high availability within AWS regions through managed infrastructure and automatic failover capabilities:

- Disaster recovery strategies typically involve replicating destination data stores (like S3 buckets or Redshift clusters) to other regions rather than replicating the Firehose streams themselves
- Cross-region disaster recovery is achieved through destination-level replication and recreating Firehose streams in alternate regions when needed

Recovery Objectives

Amazon Data Firehose helps customers meet RPO objectives through its reliable data delivery with at-least-once semantics and automatic retry mechanisms. RPO can be as low as seconds due to real-time streaming capabilities and configurable buffer intervals. RTO objectives are supported through automatic scaling, managed infrastructure, and quick stream recreation capabilities. The service's 99.9% SLA uptime guarantee and built-in error handling help maintain consistent data flow. Buffer management allows customers to balance between data freshness and delivery efficiency to meet specific RPO requirements.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/firehose/latest/dev/limits.html

FAQ: https://aws.amazon.com/firehose/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/firehose/latest/dev/index.html

User Guide: https://docs.aws.amazon.com/firehose/latest/dev/what-is-this-service.html

API Reference:
https://docs.aws.amazon.com/firehose/latest/APIReference/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/firehose/pricing/

Cost Optimization

Amazon Data Firehose offers several unique cost optimization opportunities through its tiered pricing model and pay-as-you-go structure. Optimizing record sizes to avoid the 5KB minimum billing increment for Direct PUT sources can significantly reduce costs:

- Batch processing using PutRecordBatch API is more cost-effective than individual PutRecord calls
- Buffer size and interval tuning helps optimize delivery batches, reducing per-request costs at destinations
- Format conversion to columnar formats like Parquet reduces storage costs and improves query performance
- Dynamic partitioning, while adding processing costs, can reduce downstream query costs significantly
- Compression algorithms reduce data transfer and storage costs
- Choosing appropriate data sources (Direct PUT vs
- MSK pricing tiers) based on volume can optimize ingestion costs

Savings Considerations

Primary cost savings come from eliminating infrastructure management overhead and reducing operational complexity compared to self-managed streaming solutions. No upfront costs or minimum commitments provide flexibility for variable workloads:

- Tiered pricing rewards high-volume users with lower per-GB rates as usage scales
- Automatic scaling eliminates over-provisioning costs common with traditional infrastructure
- Built-in compression and format conversion reduce downstream storage and compute costs
- Error handling and retry mechanisms reduce data loss costs and operational overhead
- Integration with managed AWS services eliminates need for custom integration development
- The serverless model means customers only pay for actual data processing rather than idle infrastructure capacity, providing significant savings for variable or unpredictable workloads

Service Name

Amazon Detective

Service Overview

Amazon Detective is a security investigation service that helps users analyze and identify the root cause of security findings or suspicious activities in their AWS environment. It automatically collects log data from various AWS resources including Amazon VPC Flow Logs, AWS CloudTrail logs, and Amazon EKS audit logs, then uses machine learning, statistical analysis, and graph theory to generate visualizations for faster and more efficient security investigations. Detective constructs a behavior graph that creates a unified view of resource behaviors and entity relationships, enabling security analysts to conduct thorough investigations with contextual insights and interactive visualizations powered by generative AI.

Key Use Cases

- Security incident investigation: Analyzing GuardDuty findings and Security Hub alerts to determine root cause and scope of security events
- Threat hunting: Proactively searching for indicators of compromise and suspicious activities across AWS environments
- Malware analysis: Investigating EC2 instances for potential malware compromises and understanding attack chains
- Compliance reporting: Supporting security compliance requirements through detailed investigation capabilities and audit trails
- Multi-account security monitoring: Centralizing security investigations across multiple AWS accounts through behavior graphs

Features

- **Behavior Graph Construction**: Creates linked datasets from log data using machine learning and graph theory to visualize entity relationships and behaviors
- **Detective Investigations**: Automated investigations using machine learning models and threat intelligence to analyze IAM users and roles for indicators of compromise
- **Finding Groups**: Consolidates related security findings and events into single navigable timelines for holistic incident analysis
- **Interactive Visualizations**: Provides guided visualizations and generative AI insights to help analysts understand attack patterns and investigate security events
- **Multi-Account Management**: Supports administrator and member account structures with AWS Organizations integration for centralized security monitoring

- **Security Lake Integration**: Enables querying and retrieval of raw log data through Amazon Security Lake integration

Value Proposition

Amazon Detective provides unique value through its automated behavior graph construction that creates a unified view of security events across AWS environments, eliminating the need for manual log correlation. The service significantly reduces investigation time by using machine learning to surface critical insights and providing guided visualizations that help analysts quickly understand attack chains:

- Detective's integration with GuardDuty, Security Hub, and Security Lake creates a seamless security workflow, while its automated investigations capability performs initial security analysis automatically
- The service requires no upfront data source integration or complex configurations, making it simple to deploy while providing enterprise-scale multi-account support through AWS Organizations integration

Service Integration

Amazon Detective deeply integrates with core AWS security services to provide comprehensive investigation capabilities. It automatically ingests findings from Amazon GuardDuty and AWS Security Hub, creating seamless investigation workflows from detection to analysis:

- Detective integrates with Amazon Security Lake for advanced querying and raw log data retrieval, while supporting AWS Organizations for multi-account security management
- The service collects data from AWS CloudTrail, Amazon VPC Flow Logs, and Amazon EKS audit logs automatically
- Additional integrations include AWS Identity and Access Management for access control, Amazon EventBridge for automation workflows, and AWS Lambda for custom investigation logic
- Detective also supports integration with third-party security tools through its APIs

Onboarding and Offboarding

Customers can get started with Amazon Detective through the AWS Management Console, AWS CLI, or Detective API by simply enabling the service in their desired AWS Region. Upon enabling, Detective automatically creates a Region-specific behavior graph with the account as administrator and begins a 30-day free trial:

- Data ingestion starts immediately from CloudTrail and VPC Flow Logs, with visualizations typically available within 2 hours

- Administrator accounts can invite member accounts to contribute data to the behavior graph for multi-account investigations
- The service requires minimal configuration as it automatically discovers and ingests relevant log data sources
- Detective works best when integrated with GuardDuty and Security Hub for comprehensive security coverage

Getting Started Guide:
https://docs.aws.amazon.com/detective/latest/userguide/detective-setup.html

Data Removal

When disabling Amazon Detective, the behavior graph and all associated data are permanently and irreversibly deleted. Customers can disable Detective through the console, Detective API, or AWS CLI. For administrator accounts, disabling Detective removes the entire behavior graph including all member account data. Member accounts can be individually removed using the DeleteMembers operation, which stops data ingestion from those accounts while preserving existing data in the behavior graph. AWS provides Python scripts for bulk account management across multiple regions during offboarding.

Backup/Restore and Disaster Recovery

Amazon Detective does not provide traditional backup and disaster recovery capabilities as it is an analytical service that processes and visualizes log data rather than storing critical business data. Detective maintains up to one year of aggregated data for analysis purposes, but this data cannot be backed up or restored by customers. The service operates as a managed service where AWS handles the underlying infrastructure resilience and data durability within each region.

Backup Capabilities

Amazon Detective does not have backup capabilities and does not integrate with AWS Backup service. The service is designed as an analytical tool that processes log data to create behavior graphs and investigations, rather than storing data that requires backup protection. Detective automatically maintains historical data for analysis purposes but does not provide customer-controlled backup or restore functionality.

Disaster Recovery

Amazon Detective does not support disaster recovery in the traditional sense and does not integrate with AWS Elastic Disaster Recovery. As a regional service, Detective operates independently in each AWS Region where it is enabled:

- Customers must enable Detective separately in each region where they require security investigation capabilities

- The service itself does not require disaster recovery planning as it is an analytical service

Recovery Objectives

Amazon Detective does not directly help customers meet RPO and RTO objectives as it is not a data protection or disaster recovery service. Detective serves as a security investigation tool that analyzes existing log data to help understand security incidents. The service can assist in post-incident analysis and forensics to understand the timeline and impact of security events, but does not provide backup or recovery capabilities.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/detective/latest/userguide/regions-limitations.html

FAQ: https://aws.amazon.com/detective/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/detective/latest/userguide/index.html

User Guide: https://docs.aws.amazon.com/detective/latest/userguide/index.html

API Reference: https://docs.aws.amazon.com/detective/latest/APIReference/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/detective/pricing/

Cost Optimization

Amazon Detective offers several cost optimization options including a 30-day free trial for new behavior graphs and tiered pricing that reduces costs as data volume increases. The first 1,000 GB costs $2.00 per GB, decreasing to $0.25 per GB for volumes over 10,000 GB monthly:

- Customers can optimize costs by carefully selecting which accounts contribute to behavior graphs and monitoring data volume through the Usage page
- The service automatically aggregates and processes data to reduce storage requirements compared to raw log retention
- Customers can also disable Detective in regions where it’s not needed and use selective member account management to control data ingestion volumes

Savings Considerations

Main cost savings with Amazon Detective come from eliminating the need for separate log analysis infrastructure and reducing investigation time through automated analysis. The service's tiered pricing model provides significant per-GB cost reductions for high-volume customers, with costs dropping by 87.5% from the lowest to highest tier:

- Detective's automated investigations and finding groups reduce the time security analysts spend on manual investigation tasks, leading to operational cost savings
- The service's integration with existing AWS security services eliminates the need for additional third-party security tools
- Organizations can achieve savings through centralized multi-account management, avoiding duplicate investigation infrastructure across business units while maintaining comprehensive security visibility

Service Name

Amazon DevOps Guru

Service Overview

Amazon DevOps Guru is a fully managed operations service that uses machine learning to analyze operational data, application metrics, and events to identify deviations from normal operating patterns. It consolidates operational data from various sources including Amazon CloudWatch metrics, AWS Config, AWS CloudFormation, and AWS X-Ray to provide a single-console dashboard for analyzing anomalies. The service generates both reactive insights to address current issues and proactive insights to prevent future operational problems, automatically providing intelligent recommendations to help developers and operators enhance application performance and availability.

Key Use Cases

- Anomaly Detection: Identify unusual application behavior such as increased latency, error rates, and resource constraints that could lead to outages
- Proactive Issue Prevention: Generate proactive insights to predict and prevent future operational problems before they impact applications
- Automated Incident Response: Integrate with PagerDuty, AWS Systems Manager Incident Manager, and third-party tools to automate incident response workflows
- Multi-account Operations: Aggregate insights across multiple AWS accounts within an organization for centralized operations management
- Cost Reduction: Reduce operational overhead by automatically detecting issues and providing contextual recommendations to minimize debugging time

Features

- **ML-Powered Anomaly Detection**: Continuously analyzes metrics, logs, events, and traces to establish normal bounds for application behavior and detect deviations
- **Reactive and Proactive Insights**: Generates insights to address current issues and predict future problems with contextual recommendations
- **Multi-Account Insight Aggregation**: Enables centralized management of insights across multiple AWS accounts within an organization
- **Log Anomaly Detection**: Analyzes CloudWatch logs to identify anomalous patterns and provide recommendations
- **Third-Party Integrations**: Integrates with PagerDuty, Atlassian Opsgenie, ServiceNow, and other incident management tools
- **Automated Alert Configuration**: Automatically configures alarms and notifications through Amazon SNS and AWS Systems Manager

Value Proposition

Amazon DevOps Guru offers significant advantages over traditional monitoring solutions by leveraging machine learning trained on years of Amazon.com and AWS operational expertise. Unlike basic alerting systems, it provides intelligent insights with contextual recommendations, reducing false positives and debugging time:

- The service requires no additional software installation or complex configuration, offering one-click deployment
- It consolidates data from multiple AWS services into a single dashboard, eliminating the need to manage multiple monitoring tools
- The pay-per-use pricing model with no upfront commitments makes it cost-effective, while automated integration with incident management tools streamlines operations workflows

Service Integration

DevOps Guru deeply integrates with core AWS services to provide comprehensive operational insights. It analyzes Amazon CloudWatch metrics and logs, AWS Config configuration data, AWS CloudFormation stack information, and AWS X-Ray traces:

- The service automatically integrates with AWS Systems Manager to create OpsItems for incident management and uses Amazon EventBridge for event-driven workflows
- It leverages Amazon SNS for notifications and works with AWS CloudTrail for audit events
- Additional integrations include Amazon ECS, EKS, Elastic Beanstalk, Lambda, S3, DynamoDB, RDS, and API Gateway
- The service also supports AWS Organizations for multi-account management and AWS Backup for operational continuity

Onboarding and Offboarding

Getting started with DevOps Guru involves three simple steps with no complex setup required. First, customers sign up for an AWS account and ensure they have appropriate IAM permissions:

- Second, they enable DevOps Guru through the AWS Management Console, CLI, or API in supported regions
- Third, they specify resource coverage by choosing to analyze all resources in their account, specific CloudFormation stacks, or resources with particular tags
- The service is fully managed and requires no software installation
- Customers can estimate costs using the AWS Pricing Calculator and begin receiving insights immediately after enabling the service

- The AWS Free Tier provides 7,200 hours of resource analysis per pricing group for three months

Getting Started Guide: https://docs.aws.amazon.com/devops-guru/latest/userguide/getting-started.html

Data Removal

To offboard from DevOps Guru and remove data, customers update their AWS analysis coverage to exclude resources from monitoring, effectively stopping new data collection and analysis. Existing insights and recommendations remain accessible through the console until manually deleted. Customers can disable the service entirely through the console or API, which stops all resource analysis and prevents new charges. All operational data analyzed by DevOps Guru is derived from existing AWS services and remains in those source services after offboarding.

Backup/Restore and Disaster Recovery

DevOps Guru itself does not provide backup and disaster recovery capabilities as it is a monitoring and analysis service. The service analyzes operational data from other AWS services but does not store application data requiring backup. DevOps Guru's insights and recommendations are ephemeral operational intelligence. For disaster recovery, customers should implement appropriate backup strategies for the underlying AWS resources that DevOps Guru monitors, such as using AWS Backup for supported services.

Backup Capabilities

DevOps Guru does not have built-in backup capabilities and does not integrate with AWS Backup. The service focuses on operational analysis and insight generation rather than data preservation:

- It analyzes metrics, logs, and events from other AWS services without storing the underlying application data
- Customers must implement backup strategies for their monitored resources using appropriate AWS backup services

Disaster Recovery

DevOps Guru does not directly support disaster recovery scenarios and does not integrate with AWS Elastic Disaster Recovery. As a monitoring service, it provides operational insights that can help identify issues affecting disaster recovery preparedness. The service itself is regionally deployed, so customers should consider multi-region monitoring strategies for comprehensive operational visibility during disaster recovery scenarios.

Recovery Objectives

DevOps Guru helps customers meet RPO and RTO objectives indirectly by providing early detection of performance anomalies and potential issues that could impact recovery processes. The service's proactive insights can identify problems before they escalate into outages, potentially preventing the need for disaster recovery activation. By integrating with incident management tools, it can accelerate response times during operational issues, contributing to faster recovery.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/devops-guru/latest/userguide/quotas.html

FAQ: https://aws.amazon.com/devops-guru/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/devops-guru/latest/userguide/welcome.html

User Guide: https://docs.aws.amazon.com/devops-guru/latest/userguide/welcome.html

API Reference: https://docs.aws.amazon.com/devops-guru/latest/APIReference/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/devops-guru/pricing/

Cost Optimization

DevOps Guru offers several unique cost optimization features including usage-based pricing with no upfront commitments or minimum fees. The service provides cost estimation tools to help customers understand expenses before enabling analysis:

- Customers pay only for active resources that produce metrics, events, or logs during each hour, not for idle resources
- The AWS Free Tier includes 7,200 hours of resource analysis per pricing group for three months
- By providing early detection of operational issues, the service can prevent costly outages and reduce the need for extensive troubleshooting resources, offering potential cost savings through improved operational efficiency

Savings Considerations

The primary cost savings from DevOps Guru come from reduced operational overhead and prevented outages rather than direct service cost reductions. The service can significantly decrease mean time to resolution (MTTR) by providing contextual insights and recommendations, reducing the need for manual investigation:

- Early detection of anomalies can prevent minor issues from escalating into costly production outages
- Integration with automated incident response tools reduces manual intervention requirements
- The machine learning-powered approach eliminates false positives common in traditional monitoring, reducing alert fatigue and operational costs
- Organizations can optimize their monitoring spend by using DevOps Guru's intelligent analysis instead of maintaining multiple separate monitoring tools

Service Name

Amazon DocumentDB (with MongoDB compatibility)

Service Overview

Amazon DocumentDB is a fully managed, MongoDB-compatible document database service provided by AWS. It enables organizations to store, query, and index JSON-like documents with complex queries, indexing, and aggregation capabilities while using existing MongoDB drivers and tools with minimal changes. The service separates storage and compute, allowing independent scaling, and provides enterprise-grade reliability with automatic backup, high availability across multiple Availability Zones, and built-in security features including VPC isolation and encryption.

Key Use Cases

- Content Management Systems: Managing user-generated content like images, comments, and videos with flexible schema requirements for various content types
- E-commerce Applications: Handling product catalogs, customer profiles, shopping carts, and order management with complex nested data structures
- Mobile and Web Applications: Storing user profiles, session data, and application metadata with variable attributes and rapid development cycles
- IoT Applications: Collecting and analyzing sensor data, device telemetry, and time-series information from connected devices
- Real-time Analytics: Processing operational data for business intelligence while maintaining concurrent access for analysis and reporting

Features

- **MongoDB Compatibility**: Compatible with MongoDB 3.6, 4.0, 5.0, and 8.0 APIs, supporting existing drivers, tools, and applications
- **Serverless Configuration**: Automatic capacity scaling with DocumentDB Capacity Units (DCUs) providing up to 90% cost savings for variable workloads
- **Elastic Clusters**: Virtually limitless scale supporting millions of reads/writes with petabytes of storage capacity across multiple shards
- **Vector Search**: HNSW and IVFFlat indexing methods for semantic search with parallel index building reducing build time by up to 30x
- **ACID Transactions**: Multi-document ACID transactions ensuring data consistency across operations
- **Point-in-Time Recovery**: Restore clusters to any second within 1-35 day backup retention period with continuous incremental backups
- **Global Clusters**: Multi-region clusters with primary and up to 5 secondary regions for disaster recovery and low-latency global reads

- **Advanced Query Capabilities**: Aggregation pipelines, geospatial querying, text search, and complex array operations

Value Proposition

Amazon DocumentDB provides superior value through its fully managed nature, eliminating database administration overhead while delivering twice the throughput of traditional managed MongoDB services. The service offers enterprise-grade reliability with six-way data replication across three Availability Zones, handling up to two data copy losses without affecting writes and three losses without affecting reads:

- Its MongoDB compatibility enables seamless migration with existing applications, drivers, and tools, while providing enhanced security through VPC isolation, encryption at rest and in transit, and IAM integration
- The serverless option delivers significant cost optimization for variable workloads, and the elastic clusters provide virtually unlimited scalability for demanding applications

Service Integration

Amazon DocumentDB integrates deeply with the AWS ecosystem through multiple core services. AWS Database Migration Service (DMS) enables seamless migration from MongoDB and other databases with minimal downtime:

- Amazon CloudWatch provides comprehensive monitoring, metrics, and alarms for performance tracking
- AWS Glue supports data cataloging and ETL operations for analytics workflows
- Amazon S3 stores automated backups and snapshots durably
- Amazon VPC ensures network isolation and security
- AWS Identity and Access Management (IAM) provides granular access control and authentication
- Integration with Amazon Redshift and Amazon OpenSearch Service enables analytics and search capabilities, while AWS Backup supports centralized backup management across AWS services

Onboarding and Offboarding

Customers begin by creating an AWS account and setting up IAM permissions, then create a DocumentDB cluster through the AWS Management Console, AWS CLI, or CloudFormation templates. The process requires configuring VPC settings with subnets across multiple Availability Zones, selecting instance types, and defining security groups:

- AWS CloudShell provides immediate connectivity for testing and initial data operations

- The getting started guide enables cluster creation and data operations in under five minutes
- Customers can migrate existing MongoDB data using AWS DMS, native MongoDB tools (mongodump/mongorestore), or third-party migration utilities
- Free tier eligibility allows new customers to explore the service with 750 hours of t3.medium instance usage monthly

Getting Started Guide:
https://docs.aws.amazon.com/documentdb/latest/developerguide/get-started-guide.html

Data Removal

Customers can remove data by deleting individual documents, collections, or entire databases using MongoDB commands. Cluster-level removal involves deleting the cluster through AWS console or CLI, which removes all instances and data. Manual snapshots can be retained or deleted separately. Automated backups are deleted when the cluster is removed, but manual snapshots persist until explicitly deleted. For complete data removal, customers must delete both the cluster and all associated manual snapshots. The deletion process is irreversible, requiring careful consideration and backup verification.

Backup/Restore and Disaster Recovery

Amazon DocumentDB provides automated continuous incremental backups to Amazon S3 with 1-35 day retention periods, enabling point-in-time recovery to any second. Manual snapshots offer long-term retention beyond the automatic backup period. Global clusters support disaster recovery with primary and secondary regions, providing cross-region data replication. The service offers near-instant crash recovery typically under 30 seconds using log-structured storage. Backup operations have zero impact on cluster performance.

Backup Capabilities

DocumentDB includes comprehensive backup capabilities with automatic continuous incremental backups stored in Amazon S3. The service supports manual cluster snapshots that can be copied across regions and shared with other AWS accounts:

- Backup storage up to 100% of cluster storage incurs no additional charges
- AWS Backup service integration enables centralized backup management and policy-based protection
- Backups are encrypted using the same KMS keys as the source cluster

Disaster Recovery

DocumentDB supports disaster recovery through Global Clusters spanning multiple AWS regions with automatic failover capabilities. The service replicates data six ways across three Availability Zones within each region, providing built-in fault tolerance:

- Cross-region snapshots enable recovery in different geographical locations
- The log-structured storage system ensures rapid recovery with minimal data loss
- AWS Elastic Disaster Recovery integration is not explicitly supported, but Global Clusters provide comprehensive disaster recovery functionality

Recovery Objectives

DocumentDB helps meet RPO objectives through continuous incremental backups with point-in-time recovery granularity to the second, minimizing potential data loss. RTO objectives are addressed through near-instant crash recovery under 30 seconds, automatic failover to read replicas within minutes, and Global Cluster failover capabilities. The six-way data replication across Availability Zones ensures high availability and rapid recovery. Customers can achieve aggressive RPO/RTO targets through proper architecture using read replicas and Global Clusters.

Service Constraints

Service Quotas:
https://docs.aws.amazon.com/documentdb/latest/developerguide/limits.html

FAQ: https://aws.amazon.com/documentdb/faqs/

Technical Requirements

Service Documentation:
https://docs.aws.amazon.com/documentdb/latest/developerguide/what-is.html

User Guide: https://docs.aws.amazon.com/documentdb/latest/developerguide/

API Reference: https://docs.aws.amazon.com/documentdb/latest/developerguide/api-reference.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/documentdb/pricing/

Cost Optimization

DocumentDB offers unique cost optimization through serverless configurations providing up to 90% savings for variable workloads compared to provisioning for peak capacity. Database Savings Plans deliver up to 35% savings with 1-year commitments:

- I/O-Optimized storage configurations provide predictable pricing for I/O-intensive applications

- Customers can optimize costs through right-sizing instances, utilizing read replicas to distribute workloads, implementing efficient indexing strategies, and leveraging storage auto-scaling
- The service offers single-instance durability options and the ability to pause/resume clusters for development environments
- Per-second billing with 10-minute minimums enables precise cost control

Savings Considerations

Primary cost savings come from the fully managed service eliminating operational overhead and administrative costs associated with self-managed MongoDB deployments. Serverless configurations automatically scale capacity based on demand, avoiding overprovisioning costs:

- Database Savings Plans provide significant discounts for predictable workloads
- The separation of compute and storage allows independent scaling, optimizing resource utilization
- Automated backup and storage management reduce operational costs
- Multi-AZ replication is included without additional charges
- Free backup storage up to 100% of cluster storage reduces backup costs
- Customers achieve additional savings through proper instance sizing, efficient query design, and leveraging read replicas for read-heavy workloads

Service Name

Amazon DynamoDB

Service Overview

Amazon DynamoDB is a fully managed, serverless NoSQL database service that delivers single-digit millisecond performance at any scale. It supports both key-value and document data models with a flexible schema, allowing applications to store and retrieve large amounts of data without infrastructure management. DynamoDB automatically handles hardware provisioning, setup, configuration, replication, cluster scaling, patching, and monitoring. The service offers built-in security features, continuous backups, multi-region replication, in-memory caching, and seamless integration with other AWS services for building modern, scalable applications.

Key Use Cases

- Adtech applications: High-performance ad serving and real-time bidding platforms requiring low latency and horizontal scaling
- Mobile and web applications: Session management, user profiles, and gaming leaderboards for applications serving millions of users
- Gaming applications: Player data, game state management, and global leaderboards with multi-region support
- Internet of Things (IoT): Time-series data collection and device management for sensor networks and IoT applications
- Real-time analytics: Event-driven architectures using DynamoDB Streams for data processing and monitoring
- Financial services: Transaction processing, fraud detection, and trading platforms requiring ACID transactions
- E-commerce and retail: Product catalogs, shopping carts, and inventory management systems
- Generative AI applications: Storage for chat history, user preferences, and contextual data for AI-powered chatbots

Features

- **Serverless Architecture**: Zero infrastructure management with automatic scaling and pay-per-use pricing
- **Single-digit Millisecond Performance**: Consistent low-latency performance at any scale with optional microsecond latency via DAX
- **Global Tables**: Multi-active, multi-region replication for globally distributed applications with 99.999% availability

- **ACID Transactions**: Support for complex business logic with atomicity, consistency, isolation, and durability
- **DynamoDB Streams**: Real-time change data capture for event-driven architectures and data processing
- **Point-in-time Recovery**: Continuous backups with recovery to any second within the last 35 days
- **Encryption**: Server-side encryption at rest and in transit with AWS KMS integration
- **Auto Scaling**: Automatic capacity adjustment based on traffic patterns for cost optimization
- **Secondary Indexes**: Global and local secondary indexes for flexible query patterns
- **PartiQL Support**: SQL-compatible query language for familiar data access patterns

Value Proposition

DynamoDB offers a unique combination of serverless convenience, predictable performance, and enterprise-grade reliability that eliminates operational overhead while delivering consistent single-digit millisecond latency. Unlike traditional databases, it provides automatic scaling without capacity planning, built-in security and compliance features, and seamless integration with AWS services:

- The service's global tables enable multi-region active-active replication with up to 99.999% availability, while features like ACID transactions, point-in-time recovery, and DynamoDB Streams support complex business requirements
- Cost optimization through on-demand pricing, multiple table classes, and automatic scaling provides better economics than managing infrastructure, making it ideal for modern applications requiring scale, performance, and reliability

Service Integration

DynamoDB integrates seamlessly with core AWS services to enable comprehensive application architectures. AWS Lambda functions can be triggered by DynamoDB Streams for real-time data processing and event-driven workflows:

- Amazon S3 supports bulk import/export of data for analytics and backup scenarios
- Amazon Kinesis Data Streams provides advanced streaming capabilities for real-time analytics
- AWS CloudWatch offers monitoring, metrics, and alarms for performance optimization
- Amazon OpenSearch Service enables full-text search and analytics through zero-ETL integration
- DynamoDB Accelerator (DAX) provides microsecond latency caching, while AWS Backup manages enterprise backup policies

- Integration with Amazon SageMaker and Amazon Redshift supports machine learning and data warehousing use cases through zero-ETL pipelines

Onboarding and Offboarding

Customers can start using DynamoDB immediately through the AWS Management Console, AWS CLI, NoSQL Workbench for DynamoDB, or AWS SDKs without any upfront setup or infrastructure provisioning. The service is included in the AWS Free Tier with 25 GB of storage, 25 read/write capacity units, and 2.5 million stream reads monthly:

- Getting started involves creating a table with a primary key, choosing between on-demand or provisioned capacity modes, and optionally configuring features like encryption, streams, or global tables
- AWS provides comprehensive tutorials, code samples, and the NoSQL Workbench tool for data modeling and application development
- Migration tools and best practices documentation help customers transition from relational databases or other NoSQL solutions

Getting Started Guide:
https://docs.aws.amazon.com/amazondynamodb/latest/developerguide/GettingStartedDynamoDB.html

Data Removal

To offboard from DynamoDB, customers can delete tables through the AWS Management Console, CLI, or APIs, which permanently removes all data. Before deletion, customers should create backups using on-demand backup or export data to Amazon S3. Point-in-time recovery backups are automatically deleted 35 days after table deletion. For global tables, customers must remove all regional replicas before deletion. AWS provides data export capabilities and migration tools to facilitate moving to alternative solutions. All associated resources like streams, DAX clusters, and backups should be cleaned up separately to avoid ongoing charges.

Backup/Restore and Disaster Recovery

DynamoDB provides comprehensive backup and disaster recovery capabilities including continuous backups with point-in-time recovery (PITR) for up to 35 days with per-second granularity, and on-demand backups for long-term retention. Global tables offer active-active multi-region replication for disaster recovery with automatic failover capabilities. AWS Backup integration enables centralized backup management with cross-account and cross-region copying. All backups are automatically encrypted and stored redundantly across multiple Availability Zones for durability.

Backup Capabilities

DynamoDB offers native backup capabilities with point-in-time recovery providing continuous backups for 1-35 days, and on-demand backups for manual snapshots. AWS Backup is fully supported, enabling centralized backup policies, scheduling, retention management, and advanced features like cross-region backup copying and lifecycle management:

- Backups are automatically encrypted and can be restored to new tables with the same configuration
- System backups are automatically created when tables are deleted

Disaster Recovery

DynamoDB supports comprehensive disaster recovery through global tables that provide active-active replication across AWS regions with automatic failover and up to 99.999% availability. Point-in-time recovery enables restoration from accidental deletions or corruptions. While AWS Elastic Disaster Recovery is designed for EC2-based workloads, DynamoDB's native global tables and backup features provide superior disaster recovery for database workloads with multi-region active-active architecture and automatic conflict resolution.

Recovery Objectives

DynamoDB helps customers achieve aggressive RTO and RPO objectives through global tables with near-zero RTO for region failures due to active-active architecture, and RPO approaching zero with eventual consistency or zero RPO with strong consistency mode. Point-in-time recovery provides RPO of up to 1 second and RTO measured in minutes for data restoration. On-demand backups offer customizable RPO based on backup frequency, while the multi-AZ architecture within regions provides sub-minute recovery from infrastructure failures.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/amazondynamodb/latest/developerguide/ServiceQuotas.html

FAQ: https://aws.amazon.com/dynamodb/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/amazondynamodb/latest/developerguide/Introduction.html

User Guide: https://docs.aws.amazon.com/amazondynamodb/latest/developerguide/

API Reference:
https://docs.aws.amazon.com/amazondynamodb/latest/APIReference/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/dynamodb/pricing/

Cost Optimization

DynamoDB offers unique cost optimization options including on-demand mode with 50% reduced pricing that scales to zero when unused, eliminating capacity planning overhead. The Standard-IA table class reduces storage costs by up to 60% for infrequently accessed data:

- Auto scaling automatically adjusts provisioned capacity based on traffic patterns, preventing over-provisioning
- Reserved capacity provides up to 76% savings for predictable workloads
- Global tables now offer up to 67% savings with simplified pricing
- Database Savings Plans provide additional discounts with usage commitments
- The service's serverless nature eliminates infrastructure costs, while features like DynamoDB Streams and built-in security reduce the need for additional services and associated costs

Savings Considerations

Major cost savings considerations include choosing the right capacity mode based on traffic patterns, with on-demand being cost-effective for unpredictable or intermittent workloads, while provisioned capacity with auto scaling optimizes costs for steady traffic. Implementing proper data modeling reduces unnecessary read/write operations and storage costs:

- Using appropriate table classes like Standard-IA for infrequently accessed data, removing unused indexes and streams, and leveraging the free tier effectively reduces costs
- Cross-region replication costs should be evaluated against availability requirements, while monitoring tools help identify optimization opportunities
- The elimination of database administration overhead, infrastructure management, and the pay-per-use model provides significant operational cost savings compared to traditional database solutions

Service Name

Amazon ECS Anywhere

Service Overview

Amazon ECS Anywhere is a feature of Amazon Elastic Container Service (ECS) that enables running and managing containerized workloads on on-premises servers, virtual machines, or other customer-managed infrastructure. It extends the familiar ECS control plane to external environments, allowing organizations to register on-premises instances as external instances in Amazon ECS clusters. The service uses the EXTERNAL launch type and requires installation of both the Amazon ECS agent and AWS Systems Manager (SSM) agent on on-premises infrastructure. ECS Anywhere provides a consistent developer experience across AWS and on-premises environments while leveraging AWS services for container orchestration, monitoring, and security.

Key Use Cases

- Use case 1: Modernizing applications - Moving legacy applications to containers while maintaining on-premises deployment for compliance, data sovereignty, or low-latency requirements
- Use case 2: Edge computing and data processing - Running containerized data-processing workloads at edge locations on customer hardware to maintain reduced latency and process data locally
- Use case 3: Hybrid cloud architectures - Creating stretched environments that span AWS and on-premises infrastructure for gradual cloud migration, capacity bursting, or maintaining capital investments

Features

- **External Instance Registration**: Register on-premises servers or VMs as external instances in ECS clusters using activation codes and installation scripts
- **Hybrid Workload Management**: Run tasks and services across AWS and on-premises infrastructure using the same ECS APIs and tooling
- **AWS Service Integration**: Leverage AWS services like IAM, CloudWatch, ECR, and Systems Manager from on-premises environments
- **Secure Communication**: Mutual TLS authentication and encrypted communication between on-premises compute and AWS cloud with automatic credential rotation
- **Multi-Platform Support**: Support for various Linux distributions, Windows Server versions, and both x86_64 and ARM64 architectures

Value Proposition

Amazon ECS Anywhere provides organizations with a unified container orchestration platform that spans AWS and on-premises environments, eliminating the need for separate container management solutions. It enables customers to maintain existing on-premises investments while benefiting from AWS's managed services, security, and tooling:

- The service offers enhanced security features, standardized deployment processes, and cost efficiency by avoiding duplicate infrastructure management overhead
- Organizations can meet compliance requirements, handle data sovereignty concerns, and maintain low-latency workloads on-premises while leveraging AWS's scale, reliability, and extensive service ecosystem for a consistent developer experience

Service Integration

Amazon ECS Anywhere integrates deeply with core AWS services to provide comprehensive container management capabilities. Key integrations include AWS Identity and Access Management (IAM) for authentication and authorization, Amazon Elastic Container Registry (ECR) for container image storage and vulnerability scanning, AWS Systems Manager for fleet management and secure communication, Amazon CloudWatch for monitoring and logging, AWS CloudFormation for infrastructure as code deployment, and Amazon Simple Queue Service (SQS) for connecting hybrid workloads. Additional integrations include AWS CodePipeline for CI/CD across environments, AWS Application Recovery Controller for disaster recovery, and various networking and storage services to support comprehensive hybrid architectures.

Onboarding and Offboarding

Getting started with Amazon ECS Anywhere involves creating an ECS cluster with external capacity providers, generating activation codes and IDs through the AWS console or CLI, and installing required software on on-premises infrastructure. Customers must install Docker, the AWS Systems Manager Agent (SSM Agent), and Amazon ECS container agent on their servers or VMs:

- The registration process uses an installation script that securely connects the external instance to the AWS ECS cluster
- Prerequisites include an active AWS account, AWS CLI configuration, appropriate IAM roles for external instances, and supported operating systems
- The setup can be automated using infrastructure as code tools like AWS CDK or CloudFormation for consistent deployments across multiple instances

Getting Started Guide:
https://docs.aws.amazon.com/AmazonECS/latest/developerguide/ecs-anywhere.html

Data Removal

Offboarding from Amazon ECS Anywhere involves deregistering external instances from ECS clusters, which stops all running tasks and removes the instances from AWS management. Customers can deregister instances through the AWS console, CLI, or APIs. The process includes stopping the ECS and SSM agents on on-premises infrastructure, removing AWS credentials and configuration files, and optionally uninstalling the agents and Docker. Since ECS Anywhere primarily manages metadata and container orchestration rather than storing customer data, the offboarding process focuses on disconnecting the management plane connection and cleaning up local agent installations.

Backup/Restore and Disaster Recovery

Amazon ECS Anywhere supports backup and disaster recovery through integration with AWS services and customer-managed solutions. Container images stored in Amazon ECR benefit from native multi-region replication and backup capabilities. Task definitions, cluster configurations, and service definitions can be backed up using infrastructure as code approaches with CloudFormation or CDK. For application data, customers can leverage AWS Backup for supported services or implement custom backup strategies. The service itself provides resilience through automatic reconnection capabilities when network connectivity is restored, and customers can implement multi-region ECS clusters for disaster recovery scenarios.

Backup Capabilities

ECS Anywhere has limited native backup capabilities, focusing primarily on configuration and metadata backup rather than comprehensive data backup. The service supports backing up task definitions, service configurations, and cluster settings through CloudFormation templates and infrastructure as code approaches:

- Container images are backed up through Amazon ECR's built-in versioning and cross-region replication features
- ECS Anywhere does not directly integrate with AWS Backup for the service itself, but applications running on external instances can utilize AWS Backup for supported data stores and volumes through the AWS APIs and agents

Disaster Recovery

ECS Anywhere supports disaster recovery scenarios through its distributed architecture and AWS service integrations. The service can automatically reconnect external instances when network connectivity is restored after outages:

- For comprehensive disaster recovery, customers can deploy multi-region ECS clusters and replicate container images across regions using Amazon ECR
- ECS Anywhere does not directly integrate with AWS Elastic Disaster Recovery, but the hybrid nature allows customers to maintain business continuity by running

workloads on-premises during AWS region issues, or failing over to AWS during on-premises outages

Recovery Objectives

Amazon ECS Anywhere helps customers achieve Recovery Point Objective (RPO) and Recovery Time Objective (RTO) targets through its hybrid architecture and AWS service integrations. The service enables backup and restore strategies with RTOs of 2-4 hours for less critical workloads, while supporting warm standby and multi-site active/active architectures for sub-hour RTOs. Container image replication through ECR and infrastructure as code deployment enable consistent recovery processes. The automatic reconnection capabilities and local execution during network partitions help maintain business continuity and reduce downtime during various failure scenarios.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/AmazonECS/latest/developerguide/service-quotas.html

FAQ: https://aws.amazon.com/ecs/anywhere/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/AmazonECS/latest/developerguide/ecs-anywhere.html

User Guide: https://docs.aws.amazon.com/AmazonECS/latest/developerguide/ecs-anywhere.html

API Reference: https://docs.aws.amazon.com/AmazonECS/latest/developerguide/ProgrammingGuide.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/ecs/anywhere/pricing/

Cost Optimization

Amazon ECS Anywhere offers unique cost optimization through its pay-per-managed-instance model at $0.01025 per hour per registered instance, with no upfront commitments or minimum fees. Organizations can optimize costs by leveraging existing on-premises hardware investments while avoiding duplicate infrastructure management overhead:

- The service allows selective workload placement, keeping data-intensive or latency-sensitive applications on-premises while utilizing AWS for scalable components
- Cost optimization strategies include right-sizing external instances, using spot instances in AWS for burst capacity, optimizing data transfer between regions, and implementing efficient container packing to maximize resource utilization across hybrid environments

Savings Considerations

Main cost savings with ECS Anywhere include eliminating the need for separate container orchestration platforms and reducing operational overhead through unified management. Organizations save by extending the life of existing hardware investments while gradually migrating to cloud, avoiding large-scale infrastructure replacement costs:

- The service reduces staffing costs through standardized tooling and processes across environments
- Additional savings come from optimized data transfer costs by processing data locally and only transferring results, reduced licensing costs compared to traditional enterprise container platforms, and efficient resource utilization through AWS-native scaling and management capabilities that eliminate over-provisioning typical in on-premises deployments

Service Name

Amazon EKS Anywhere

Service Overview

Amazon EKS Anywhere is a deployment option for Amazon EKS that enables customers to create and operate Kubernetes clusters on their own infrastructure. It is a customer-managed product built on the Kubernetes sub-project Cluster API (CAPI) that simplifies Kubernetes cluster management through automation of infrastructure setup and cluster lifecycle operations in on-premises and edge environments. Unlike Amazon EKS, customers are responsible for cluster lifecycle operations and maintenance. EKS Anywhere uses Amazon EKS Distro, the same Kubernetes distribution deployed by Amazon EKS, providing consistent tooling and software updates. It can run in air-gapped environments with optional AWS service integrations for observability and identity management.

Key Use Cases

- Deploying across hybrid environments: Running Kubernetes workloads consistently across on-premises data centers and AWS cloud environments
- Deploying high-performing large language models (LLMs): Supporting AI/ML workloads that require on-premises compute resources
- Maintaining data sovereignty: Keeping sensitive data and workloads within specific geographic boundaries or regulatory jurisdictions
- Air-gapped environments: Operating Kubernetes clusters in secure, disconnected environments without internet connectivity
- Edge computing: Managing containerized applications at edge locations with limited connectivity to cloud services

Features

- **Infrastructure Support**: Supports various infrastructure types including VMware vSphere, bare metal, Nutanix, Apache CloudStack, and AWS Snow
- **Operating System Options**: Provides Bottlerocket as default OS with Ubuntu and Red Hat Enterprise Linux (RHEL) alternatives
- **Air-gapped Operation**: Can run disconnected from AWS Regions in secure, isolated environments
- **AWS Service Integration**: Optional integrations with AWS IAM, Certificate Manager, CloudWatch Logs, and Managed Prometheus
- **EKS Connector**: Software agent enabling cluster registration and management through Amazon EKS console

- **Curated Packages**: AWS-vended software add-ons for load balancing, observability, and auto-scaling capabilities
- **Extended Kubernetes Support**: Provides extended support for Kubernetes versions after community support expires

Value Proposition

Amazon EKS Anywhere eliminates the complexity of managing multiple third-party tools and separate support contracts required for on-premises Kubernetes deployments. It provides consistent operational tooling with Amazon EKS in the cloud, using the same Kubernetes distribution (EKS Distro) with latest security patches and updates:

- Customers can consolidate support costs through AWS support agreements, reduce operational overhead through automation of undifferentiated heavy lifting, and maintain consistency between cloud and on-premises environments
- The service offers flexibility to run in various infrastructure environments while providing optional AWS service integrations for enhanced capabilities
- This approach accelerates implementation time and reduces the fragmentation typically associated with multi-vendor Kubernetes solutions

Service Integration

Amazon EKS Anywhere integrates with several core AWS services through optional connections when not running in air-gapped mode. Key integrations include AWS Identity and Access Management (IAM) for cluster authentication and IAM Roles for Service Accounts (IRSA), Amazon Certificate Manager for certificate management, Amazon CloudWatch Logs for centralized logging via FluentBit, Amazon Managed Service for Prometheus for metrics collection using AWS Distro for OpenTelemetry, and Amazon EKS console for cluster visualization and management through the EKS Connector:

- The service also leverages Amazon EKS Distro for consistent Kubernetes distribution, Amazon S3 and Amazon ECR for secure access to releases and container images, and AWS License Manager for subscription management
- These integrations enable unified management and observability across hybrid environments

Onboarding and Offboarding

Customers begin by downloading and running the EKS Anywhere installer from the official documentation. The process involves three main steps: installing the eksctl-anywhere CLI tool, creating cluster configurations for their chosen infrastructure provider (vSphere, bare metal, etc.), and deploying clusters using the installer:

- After initial deployment, customers can optionally install the EKS Connector to view and manage clusters through the Amazon EKS console

- For enhanced support and access to curated packages, customers can purchase Amazon EKS Anywhere Enterprise Subscriptions through the AWS console, which requires AWS Enterprise Support or Enterprise On-Ramp Support plans
- The service provides integrated tooling for ongoing cluster inspection, upgrades, and scaling operations

Getting Started Guide: https://anywhere.eks.amazonaws.com/docs/getting-started/install/

Data Removal

Amazon EKS Anywhere is a customer-managed service where customers maintain full control over their data and cluster infrastructure. Data removal involves deleting cluster resources, persistent volumes, and associated storage according to standard Kubernetes deletion procedures. Since clusters run on customer-owned infrastructure, customers are responsible for securely wiping underlying storage systems and ensuring complete data removal from their hardware. Subscription cancellation can be done through the AWS console within 7 days without charge, or by contacting AWS account teams thereafter.

Backup/Restore and Disaster Recovery

Amazon EKS Anywhere relies on customer-implemented backup and disaster recovery solutions since it operates on customer-managed infrastructure. Customers can implement standard Kubernetes backup tools like Velero or other third-party solutions for cluster state and persistent volume backups. The service itself does not provide built-in backup capabilities, requiring customers to design and implement their own backup strategies using available Kubernetes ecosystem tools and infrastructure-specific backup solutions.

Backup Capabilities

Amazon EKS Anywhere does not provide native backup capabilities as it is a customer-managed service. Customers must implement their own backup solutions using standard Kubernetes tools and ecosystem solutions. The service does not integrate directly with AWS Backup, requiring customers to use third-party backup tools like Velero, or infrastructure-specific backup solutions depending on their underlying platform (vSphere, bare metal, etc.).

Disaster Recovery

Amazon EKS Anywhere does not provide built-in disaster recovery capabilities or direct integration with AWS Elastic Disaster Recovery. Customers are responsible for implementing their own disaster recovery strategies using standard Kubernetes patterns like multi-cluster deployments, data replication, and infrastructure-level disaster recovery solutions. Recovery planning must account for the underlying infrastructure platform and customer-specific requirements.

Recovery Objectives

Amazon EKS Anywhere supports customer-defined RPO and RTO objectives through architectural flexibility rather than built-in features. Customers can achieve specific recovery objectives by implementing multi-cluster deployments, configuring appropriate data replication strategies, and designing infrastructure redundancy. The service's support for various infrastructure types enables customers to design geographically distributed deployments that align with their recovery requirements and business continuity needs.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/general/latest/gr/eks.html

FAQ: https://aws.amazon.com/eks/eks-anywhere/faqs/

Technical Requirements

Service Documentation: https://anywhere.eks.amazonaws.com/docs/

User Guide: https://anywhere.eks.amazonaws.com/docs/getting-started/install/

API Reference: https://docs.aws.amazon.com/eks/latest/APIReference/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/eks/eks-anywhere/pricing/

Cost Optimization

Amazon EKS Anywhere offers several cost optimization opportunities through its flexible deployment model. The core software is available as open source at no cost, with optional Enterprise Subscriptions providing additional value through support and curated packages:

- Customers can optimize infrastructure costs by rightsizing clusters to actual workload requirements and leveraging existing on-premises hardware investments
- The flat per-cluster pricing model ($24,000 annually for 1-year terms, $18,000 annually for 3-year terms) doesn't scale with cluster size, making it cost-effective for larger deployments
- Multi-year subscriptions provide 25% cost savings compared to annual terms
- Customers can also optimize by sharing curated package access across AWS accounts within their organization and implementing efficient resource utilization strategies across their infrastructure

Savings Considerations

Primary cost savings come from leveraging existing on-premises infrastructure investments rather than migrating everything to cloud-based solutions. The 3-year subscription term provides 25% savings over 1-year terms, while the flat per-cluster pricing model benefits larger cluster deployments:

- Organizations can consolidate support costs by replacing multiple vendor contracts with unified AWS support
- Operational efficiency gains through consistent tooling and automation reduce administrative overhead costs
- By running workloads on-premises, customers can avoid cloud egress charges and optimize for workloads with predictable, long-term compute requirements
- The ability to share curated packages across AWS accounts enables cost sharing within large organizations, further reducing per-application costs

Service Name

Amazon ElastiCache

Service Overview

Amazon ElastiCache is a fully managed in-memory data store service that simplifies the deployment, operation, and scaling of distributed in-memory caches in the cloud. It supports three open-source engines: Valkey, Memcached, and Redis OSS. ElastiCache provides high performance, low latency data access by caching frequently accessed data in memory, reducing the need for frequent database reads. The service offers both serverless and node-based deployment options, with automatic failover, replication, and scaling capabilities for high availability and durability.

Key Use Cases

- Use case 1: Session caching for web applications to store user session data with high performance and low latency access
- Use case 2: Database query result caching to reduce database load and improve application response times
- Use case 3: Real-time analytics and leaderboards for gaming applications requiring sub-millisecond response times
- Use case 4: Generative AI applications for semantic caching and memory mechanisms for chatbots and AI assistants
- Use case 5: Message queuing and pub/sub systems for high-throughput, real-time communication
- Use case 6: Rate limiting and API throttling to control application traffic and prevent abuse

Features

- **Multi-Engine Support**: Support for Valkey, Memcached, and Redis OSS engines with full compatibility
- **ElastiCache Serverless**: Fully managed serverless caching that automatically scales without capacity planning
- **Multi-AZ with Automatic Failover**: High availability across multiple Availability Zones with automatic failover capability
- **Read Replicas**: Up to 5 read replicas per primary node for improved read performance and data protection
- **Global Datastore**: Cross-region replication for disaster recovery and global data access
- **Encryption**: Encryption at rest and in transit with AWS KMS integration for data security

- **Backup and Restore**: Automatic and manual snapshots with point-in-time recovery capabilities
- **VPC Support**: Network isolation within Amazon VPC with security group controls
- **Data Tiering**: Cost optimization using SSDs for least frequently used data
- **Vector Search**: Low-latency vector search capabilities for AI and machine learning applications

Value Proposition

Amazon ElastiCache offers compelling advantages over self-managed or alternative caching solutions. It eliminates the operational overhead of managing in-memory data stores, providing fully managed infrastructure with automatic patching, monitoring, and scaling:

- ElastiCache delivers sub-millisecond response times with high throughput, supporting millions of operations per second
- The service provides enterprise-grade security with encryption, VPC isolation, and compliance certifications including HIPAA, PCI DSS, and FedRAMP
- Cost optimization features include reserved instances, serverless pricing, and data tiering
- ElastiCache's seamless AWS integration, 99.99% availability SLA, and comprehensive monitoring make it ideal for mission-critical applications requiring reliable, high-performance caching

Service Integration

Amazon ElastiCache integrates seamlessly with core AWS services to provide comprehensive application architectures. It works as a caching layer for Amazon RDS and Amazon DynamoDB to improve database performance:

- Integration with Amazon EC2 and AWS Lambda enables application-level caching for compute workloads
- ElastiCache connects with Amazon VPC for network security and isolation, AWS IAM for access control, and AWS KMS for encryption key management
- It integrates with Amazon CloudWatch for monitoring and alerting, AWS CloudFormation for infrastructure as code deployment, and AWS Backup for centralized backup management
- The service also supports Amazon SNS for event notifications and AWS Auto Scaling for dynamic capacity adjustment

Onboarding and Offboarding

Getting started with Amazon ElastiCache involves several straightforward steps. Customers begin by setting up an AWS account and configuring IAM permissions for ElastiCache access:

- They can choose between ElastiCache Serverless for automatic scaling or node-based clusters for more control
- The AWS Management Console, AWS CLI, or APIs can create clusters by selecting the engine (Valkey, Memcached, or Redis OSS), configuring VPC settings, and setting up security groups
- ElastiCache provides getting started tutorials for Python integration and Lambda configuration
- The service includes comprehensive documentation, best practices guides, and sample code to accelerate implementation
- AWS Free Tier offers 750 hours of cache.t3.micro node usage for new users

Getting Started Guide:
https://docs.aws.amazon.com/AmazonElastiCache/latest/dg/GettingStarted.html

Data Removal

ElastiCache offboarding involves deleting clusters through the AWS Management Console, CLI, or API. For node-based clusters, customers can take final backups before deletion to preserve data. When deleting replication groups, ElastiCache automatically takes a final backup from the primary node. Serverless caches can be deleted immediately without backup requirements. All associated resources including subnet groups, parameter groups, and security groups should be cleaned up. Data is permanently removed from memory upon cluster deletion, and any stored backups in S3 must be manually deleted to avoid ongoing storage charges.

Backup/Restore and Disaster Recovery

ElastiCache provides comprehensive backup and disaster recovery capabilities through automatic and manual snapshots stored in Amazon S3. Backups are supported for Valkey, Redis OSS, and Serverless Memcached. Automatic backups can be scheduled with configurable retention periods, while manual backups are retained indefinitely. Cross-region disaster recovery is available through Global Datastore, enabling replication across multiple regions. Multi-AZ deployment with automatic failover provides high availability within regions. Backups include all cluster data and metadata, supporting point-in-time recovery and cluster seeding for new deployments.

Backup Capabilities

ElastiCache has extensive backup capabilities including automatic daily snapshots and on-demand manual backups. Backups are stored in Amazon S3 with configurable retention policies:

- The service supports up to 24 manual backups per day per serverless cache and 20 manual backups per node-based cluster
- However, ElastiCache does not directly integrate with AWS Backup service
- Backup creation can impact performance on node-based clusters, so creating backups from read replicas is recommended
- Backups can be exported to S3 for analysis and restored to new clusters for disaster recovery or development purposes

Disaster Recovery

ElastiCache supports disaster recovery through multiple mechanisms but does not integrate with AWS Elastic Disaster Recovery. The service offers Global Datastore for cross-region replication, enabling automatic failover to secondary regions:

- Multi-AZ deployments provide automatic failover within regions for high availability
- Cross-region backup copying enables data protection across regions
- ElastiCache's disaster recovery capabilities include read replicas, automatic failover detection, and the ability to promote read replicas to primary nodes
- These features provide comprehensive disaster recovery without requiring AWS Elastic Disaster Recovery integration

Recovery Objectives

ElastiCache helps customers meet RPO and RTO objectives through multiple mechanisms. Global Datastore provides near real-time cross-region replication with RPO typically under 1 second and RTO of minutes. Multi-AZ deployments offer automatic failover with RTO typically under 1 minute and minimal data loss. Backup and restore capabilities support longer-term recovery scenarios with RPO determined by backup frequency. ElastiCache Serverless provides built-in high availability with automatic scaling and fault tolerance. Read replicas enable quick promotion for disaster recovery scenarios, and the service's monitoring capabilities help detect issues proactively to minimize downtime.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/general/latest/gr/elasticache-service.html

FAQ: https://aws.amazon.com/elasticache/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/AmazonElastiCache/latest/dg/index.html

User Guide: https://docs.aws.amazon.com/AmazonElastiCache/latest/dg/

API Reference: https://docs.aws.amazon.com/AmazonElastiCache/latest/APIReference/

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/elasticache/pricing/

Cost Optimization

ElastiCache offers several unique cost optimization features including ElastiCache Serverless with pay-per-use pricing based on data storage (GB-hours) and compute (ECPUs). Reserved instances provide up to 72% savings with 1-3 year commitments:

- Data tiering using R6gd nodes moves least frequently used data from memory to SSD storage for cost reduction
- Database Savings Plans offer up to 35% discounts with 1-year commitments
- The Valkey engine provides 20% lower pricing for self-designed clusters and 33% lower for serverless compared to other engines
- Rightsizing tools and monitoring help identify underutilized resources, and the AWS Free Tier provides 750 hours monthly for testing and development

Savings Considerations

Key cost savings considerations include choosing appropriate deployment models (serverless vs. node-based) based on usage patterns:

- Reserved instances offer significant discounts for predictable workloads, while on-demand pricing suits variable usage
- Data tiering reduces memory costs by utilizing SSDs for infrequently accessed data
- Cross-region data transfer costs should be considered when implementing Global Datastore
- Monitoring and rightsizing prevent over-provisioning, while automated scaling helps match capacity to demand
- Upgrading to Valkey engine provides immediate cost reductions
- Backup storage optimization and proper retention policies minimize S3 costs
- Regular cost reviews using AWS Cost Explorer and CUDOS dashboard help identify additional optimization opportunities for sustained savings

Service Name

Amazon Elastic Block Store (EBS)

Service Overview

Amazon Elastic Block Store (EBS) provides scalable, high-performance block storage resources designed for use with Amazon Elastic Compute Cloud (EC2) instances. EBS volumes function as raw, unformatted block devices that can be attached to EC2 instances and used like traditional hard drives for storing files, installing applications, or hosting databases. EBS offers multiple volume types optimized for different workloads, ranging from cost-effective general-purpose storage to high-performance provisioned IOPS volumes. The service includes point-in-time snapshots for backup and recovery, encryption capabilities, and features like elastic volumes that allow dynamic scaling of capacity and performance without downtime.

Key Use Cases

- Database storage: Primary storage for relational and NoSQL databases like MySQL, PostgreSQL, MongoDB, and enterprise databases such as SAP HANA, Microsoft SQL Server, and Oracle
- File systems and application storage: Boot volumes for EC2 instances, application data storage, and shared storage for distributed applications
- Big data and analytics: Storage for data warehouses, Hadoop analytics clusters, log processing, and MapReduce workloads
- Virtual desktops and development environments: Persistent storage for VDI deployments, development and testing environments
- Backup and disaster recovery: Snapshot-based backup solutions and cross-region data replication for business continuity

Features

- **Multiple Volume Types**: Six volume types including General Purpose SSD (gp3, gp2), Provisioned IOPS SSD (io2 Block Express, io1), and HDD (st1, sc1) optimized for different performance and cost requirements
- **EBS Snapshots**: Point-in-time incremental backups stored in Amazon S3 that can be used to create new volumes, migrate data across regions, or restore data
- **Elastic Volumes**: Ability to dynamically modify volume size, performance, and type in production without service interruption
- **Encryption**: Data-at-rest and data-in-transit encryption using AWS KMS with AES-256 encryption, supporting both AWS managed and customer managed keys
- **Multi-Attach**: Attach a single EBS volume to multiple EC2 instances simultaneously within the same Availability Zone for shared storage scenarios

- **Fast Snapshot Restore**: Enables instant restoration of data from snapshots without initialization latency penalty

Value Proposition

Amazon EBS delivers superior block storage capabilities with 99.8% to 99.999% durability and the flexibility to scale performance independently from storage capacity. Unlike instance store volumes that are ephemeral, EBS provides persistent storage that survives instance failures and can be detached and reattached as needed:

- The service offers predictable performance with options to provision specific IOPS and throughput levels, comprehensive encryption capabilities, and automated backup solutions
- EBS integrates seamlessly with other AWS services and provides cost optimization through multiple volume types, allowing customers to match storage characteristics to workload requirements while maintaining high availability and data protection

Service Integration

Amazon EBS integrates deeply with Amazon EC2 as its primary compute platform, serving as persistent block storage for instances. The service leverages AWS Key Management Service (KMS) for encryption key management and integrates with Amazon Data Lifecycle Manager for automated snapshot scheduling and retention policies:

- AWS Backup provides centralized backup management across EBS volumes, while Amazon CloudWatch delivers monitoring and metrics
- EBS works with AWS CloudFormation for infrastructure as code deployments and integrates with AWS Identity and Access Management (IAM) for access control
- The service also connects with AWS Storage Gateway for hybrid deployments and AWS Elastic Disaster Recovery for comprehensive disaster recovery solutions

Onboarding and Offboarding

Getting started with Amazon EBS involves creating volumes through the EC2 console, AWS CLI, or APIs, then attaching them to EC2 instances within the same Availability Zone. Customers can create volumes from scratch or from existing snapshots, choosing the appropriate volume type based on performance requirements:

- The service supports multiple management interfaces including the AWS Management Console, AWS CLI, AWS Tools for PowerShell, and CloudFormation templates
- Initial setup includes configuring encryption settings, establishing backup policies through snapshots or AWS Backup, and implementing appropriate IAM permissions

- Best practices include right-sizing volumes for workloads, implementing regular snapshot schedules, and monitoring performance metrics through CloudWatch

Getting Started Guide: https://docs.aws.amazon.com/ebs/latest/userguide/ebs-getting-started.html

Data Removal

Data removal from EBS involves detaching volumes from EC2 instances and permanently deleting them, which cannot be undone. Before deletion, customers should create final snapshots for data retention if needed. For encrypted volumes, AWS automatically handles encryption key cleanup. Snapshots can be deleted independently, with charges stopping immediately upon deletion. The service supports secure data sanitization procedures as volumes are presented as raw block devices, allowing customers to perform data wiping operations before deletion if required for compliance purposes.

Backup/Restore and Disaster Recovery

EBS provides comprehensive backup capabilities through snapshots, which are incremental, point-in-time copies stored in Amazon S3 with cross-AZ replication. Amazon Data Lifecycle Manager automates snapshot creation, retention, and deletion policies. AWS Backup offers centralized backup management with cross-region copy capabilities. For disaster recovery, snapshots can be copied across regions, and Fast Snapshot Restore eliminates initialization latency. The service integrates with AWS Elastic Disaster Recovery for complete application-level disaster recovery solutions.

Backup Capabilities

Amazon EBS includes native snapshot functionality for point-in-time backups stored in Amazon S3. The service fully integrates with AWS Backup, providing centralized backup management, automated scheduling, retention policies, and compliance reporting:

- Snapshots are incremental, storing only changed blocks since the last snapshot, and can be automated using Amazon Data Lifecycle Manager
- Cross-region snapshot copying enables geographic backup distribution for enhanced data protection and disaster recovery scenarios

Disaster Recovery

EBS supports disaster recovery through cross-region snapshot replication and integration with AWS Elastic Disaster Recovery (DRS). Snapshots can be copied to multiple regions for geographic redundancy, and Fast Snapshot Restore enables rapid volume recreation without performance penalties:

- The service works with AWS DRS for continuous block-level replication of entire instances including EBS volumes

- Multi-Attach capability supports shared storage scenarios for high-availability architectures within the same Availability Zone

Recovery Objectives

EBS helps achieve aggressive RPO and RTO objectives through frequent automated snapshots (potentially every 15 minutes via Data Lifecycle Manager), Fast Snapshot Restore for instant volume initialization, and cross-region replication capabilities. For RPO, continuous snapshots can minimize data loss to minutes, while Fast Snapshot Restore and pre-warmed volumes can reduce RTO to minutes for volume restoration. Integration with AWS Elastic Disaster Recovery enables RPO of seconds and RTO of minutes for complete application recovery scenarios.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/ebs/latest/userguide/ebs-resource-quotas.html

FAQ: https://aws.amazon.com/ebs/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/ebs/latest/userguide/index.html

User Guide: https://docs.aws.amazon.com/ebs/latest/userguide/what-is-ebs.html

API Reference: https://docs.aws.amazon.com/AWSEC2/latest/APIReference/API_Operations_Amazon_Elastic_Block_Store.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/ebs/pricing/

Cost Optimization

EBS offers several unique cost optimization strategies including migration from gp2 to gp3 volumes for up to 20% cost savings while improving performance, right-sizing volumes to match actual usage patterns, and implementing lifecycle policies for snapshot management to reduce long-term storage costs. The service provides EBS Snapshot Archive for 75% cost reduction on long-term snapshot storage, volume cloning for efficient testing environments, and the ability to independently scale IOPS and throughput on gp3 and io2 volumes to avoid over-provisioning. Customers can optimize costs by deleting unattached volumes, using appropriate volume types for workload characteristics, and leveraging Reserved Instances for predictable workloads.

Savings Considerations

Primary cost savings opportunities include migrating legacy gp2 volumes to gp3 for immediate cost reductions and performance improvements, implementing automated snapshot lifecycle management to prevent indefinite snapshot accumulation, and regularly auditing for unattached or unused volumes. Right-sizing volumes based on actual utilization patterns rather than peak estimates, using HDD volumes (st1, sc1) for throughput-intensive workloads where IOPS are less critical, and implementing EBS Snapshot Archive for long-term retention requirements can significantly reduce costs. Monitoring CloudWatch metrics helps identify over-provisioned volumes, while using volume clones instead of full snapshots for development and testing environments provides additional savings opportunities.

Service Name

Amazon Elastic Compute Cloud (EC2)

Service Overview

Amazon Elastic Compute Cloud (Amazon EC2) is a web service that provides secure, resizable compute capacity in the cloud. It enables users to launch and manage virtual servers, called EC2 instances, in the AWS Cloud with complete control over the operating system and applications. EC2 offers various instance types with different balances of compute, memory, storage, and networking resources to meet diverse workload requirements. Users can configure security, networking, and storage for their instances, scale capacity up or down based on demand, and pay only for the resources they use. EC2 provides the foundation for many cloud computing solutions and serves as a core building block for AWS infrastructure.

Key Use Cases

- Web servers and application hosting: Running web applications, APIs, and microservices with scalable compute capacity
- High-performance computing (HPC): Scientific computing, financial modeling, batch processing, and simulation workloads requiring powerful processors
- Enterprise applications: Business-critical applications, databases, ERP systems, and development environments with customizable configurations
- Big data and analytics: Data processing, machine learning training, real-time analytics, and distributed computing workloads
- Gaming and media: Game servers, video encoding, rendering workloads, and content delivery applications
- Disaster recovery and backup: Creating backup systems, disaster recovery environments, and failover infrastructure

Features

- **Instance Types**: Over 750 instance types across 7 categories (General Purpose, Compute Optimized, Memory Optimized, Storage Optimized, Accelerated Computing, High Performance Computing, Previous Generation) with various processor, memory, storage, and networking configurations
- **Amazon Machine Images (AMIs)**: Pre-configured templates containing operating systems, applications, and software configurations for quick instance deployment
- **Elastic Block Store (EBS)**: High-performance block storage volumes that persist independently of instance lifecycle, with multiple volume types and encryption capabilities

- **Auto Scaling**: Automatically adjusts the number of instances based on demand, maintaining application availability while optimizing costs
- **Security Groups**: Virtual firewalls that control inbound and outbound traffic to instances at the protocol and port access level
- **Elastic Load Balancing**: Distributes incoming traffic across multiple instances to ensure high availability and fault tolerance
- **Spot Instances**: Access to spare EC2 capacity at up to 90% discount compared to On-Demand prices for fault-tolerant and flexible workloads
- **Enhanced Networking**: High packet-per-second performance, low latency, and high throughput networking capabilities with SR-IOV and Elastic Fabric Adapter support

Value Proposition

EC2 provides unmatched flexibility and control over computing resources in the cloud, allowing customers to choose from the broadest selection of processors (Intel, AMD, AWS Graviton, NVIDIA) and instance types to optimize performance and cost for specific workloads. The service offers multiple purchasing options including On-Demand, Reserved Instances, Savings Plans, and Spot Instances, enabling up to 90% cost savings:

- EC2's global infrastructure spans multiple regions and availability zones, ensuring high availability and fault tolerance
- Integration with 200+ AWS services provides a comprehensive cloud ecosystem, while features like Auto Scaling and Elastic Load Balancing ensure applications remain available and performant under varying loads
- The pay-as-you-go pricing model eliminates upfront investments and capacity planning guesswork

Service Integration

EC2 integrates seamlessly with core AWS services to provide comprehensive cloud solutions. Amazon VPC provides isolated networking environments with custom IP ranges, subnets, and security controls:

- Amazon EBS offers persistent block storage with multiple performance tiers and encryption
- Amazon CloudWatch monitors instance performance, health, and resource utilization with custom metrics and alarms
- AWS IAM manages access control and permissions for EC2 resources
- Elastic Load Balancing distributes traffic across instances for high availability
- Amazon Route 53 provides DNS services for domain management
- AWS Systems Manager enables instance management, patching, and configuration at scale

- Integration with Amazon S3 provides object storage for backups and data archiving, while Amazon RDS offers managed database services that work with EC2-hosted applications

Onboarding and Offboarding

Customers can get started with EC2 through the AWS Management Console using the launch instance wizard, which provides default configurations for quick deployment. The process involves selecting an Amazon Machine Image (AMI), choosing an instance type, configuring security groups and key pairs, and setting up storage:

- AWS offers a Free Tier that includes 750 hours per month of t2.micro instances for eligible customers
- The getting started tutorial guides users through launching, connecting to, and managing their first instance
- Multiple access methods are available including the console, AWS CLI, SDKs, CloudFormation, and APIs
- AWS provides extensive documentation, tutorials, and hands-on labs to help customers learn EC2 fundamentals
- For production workloads, customers should consider launch templates for consistent deployments and Auto Scaling for high availability

Getting Started Guide:
https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/EC2_GetStarted.html

Data Removal

To remove data when offboarding from EC2, customers must terminate instances, which permanently deletes all data on instance store volumes and detaches EBS volumes. EBS volumes can be separately deleted to remove persistent data. Snapshots stored in Amazon S3 should be explicitly deleted. AMIs and associated snapshots created by customers need manual deletion. Key pairs can be deleted from the console. Security groups and other network resources should be cleaned up. AWS provides CloudFormation templates to automate resource cleanup and ensure complete removal of infrastructure and associated data.

Backup/Restore and Disaster Recovery

EC2 provides comprehensive backup and disaster recovery capabilities through EBS snapshots, AMIs, and AWS Backup integration. EBS snapshots create point-in-time backups stored in Amazon S3 with 99.999999999% durability. AMIs capture complete instance configurations including operating system and applications. AWS Backup offers centralized backup management across EC2 instances with automated scheduling, lifecycle management, and cross-region replication. Instance store volumes require

application-level backup to durable storage like S3 or EBS. Multi-AZ deployments and Auto Scaling groups provide built-in fault tolerance and automatic recovery capabilities.

Backup Capabilities

EC2 offers native backup capabilities through EBS snapshots and AMIs for creating point-in-time backups. AWS Backup provides centralized backup management for EC2 instances with automated scheduling, retention policies, and compliance reporting:

- Backup jobs can be configured at the instance level with custom backup plans
- Cross-region backup replication is supported for disaster recovery scenarios
- AWS Backup enables item-level search and recovery from backups, allowing restoration of specific files without full instance recovery

Disaster Recovery

EC2 supports disaster recovery through multi-region deployments, cross-AZ redundancy, and integration with AWS Elastic Disaster Recovery. Customers can replicate instances and data across regions for geographic separation:

- Auto Scaling groups automatically replace failed instances
- EBS snapshots can be copied across regions for data recovery
- AWS Elastic Disaster Recovery provides continuous replication and fast recovery of physical, virtual, and cloud-based servers
- CloudFormation templates enable consistent infrastructure recreation in disaster scenarios

Recovery Objectives

EC2 helps customers achieve RPO and RTO objectives through multiple mechanisms. EBS snapshots can be taken as frequently as every hour for low RPO requirements. AMIs enable rapid instance recreation for improved RTO. Auto Scaling groups provide automatic instance replacement within minutes. Cross-region replication and AWS Elastic Disaster Recovery support enterprise RPO/RTO requirements. Reserved capacity and placement groups ensure resources are available during recovery scenarios. Automated backup scheduling through AWS Backup maintains consistent recovery points across the infrastructure.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/ec2-resource-limits.html

FAQ: https://aws.amazon.com/ec2/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/concepts.html

User Guide: https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/

API Reference: https://docs.aws.amazon.com/AWSEC2/latest/APIReference/

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/ec2/pricing/

Cost Optimization

EC2 offers multiple unique cost optimization options including Spot Instances providing up to 90% savings for fault-tolerant workloads, Savings Plans offering up to 72% discount for consistent usage commitments, and Reserved Instances for predictable workloads. AWS Graviton-based instances deliver up to 40% better price performance:

- Flexible instance types like M7i-flex provide optimized pricing for general-purpose workloads
- EC2 Instance Type Finder and AWS Compute Optimizer recommend optimal instance types
- Auto Scaling automatically adjusts capacity to match demand
- Hibernation allows pausing instances while preserving memory state
- Dedicated Hosts enable license mobility for cost savings
- Mixed instance types in Auto Scaling groups optimize costs while maintaining performance

Savings Considerations

Key cost savings considerations include analyzing workload patterns to choose appropriate purchasing options - use On-Demand for variable workloads, Reserved Instances for steady-state usage, and Spot Instances for fault-tolerant applications. Right-sizing instances based on actual usage prevents over-provisioning:

- Implementing Auto Scaling ensures paying only for needed capacity
- Choosing the optimal instance family and generation impacts price performance
- AWS Graviton instances offer significant cost advantages
- Storage optimization through appropriate EBS volume types and snapshot lifecycle management reduces costs

- Using placement groups and Enhanced Networking features can improve efficiency without additional charges
- Regular cost monitoring through Cost Explorer and Trusted Advisor identifies optimization opportunities

Service Name

Amazon Elastic Container Registry (ECR)

Service Overview

Amazon Elastic Container Registry (ECR) is an AWS managed container image registry service that provides secure, scalable, and reliable storage for Docker images and Open Container Initiative (OCI) compatible artifacts. ECR supports both private repositories with resource-based permissions using AWS IAM for controlled access, and public repositories for worldwide distribution. Users can push, pull, and manage container images using preferred CLI tools, with deep integration into AWS container orchestration services like Amazon ECS, Amazon EKS, and AWS Fargate for streamlined development-to-production workflows.

Key Use Cases

- Container image storage and distribution: Centralized repository for storing and distributing Docker container images within organizations or for public consumption
- Continuous integration and deployment (CI/CD): Integrated automated building, testing, and deployment of containerized applications with various CI/CD tools
- Microservices architecture: Supporting containerized microservices deployments across AWS container orchestration platforms
- Hybrid and multi-cloud deployments: Container image management across different deployment environments
- Globally distributed applications: Cross-region replication for consistent global deployments with reduced latency
- Security and compliance: Image vulnerability scanning, access control, and managed signing for enhanced container security

Features

- **Lifecycle Policies**: Automated cleanup of unused images based on defined rules for image age or tag criteria
- **Image Scanning**: Vulnerability scanning for container images with scan-on-push capability using Clair CVE database
- **Cross-Region Replication**: Automatic replication of images across AWS regions for high availability and reduced latency
- **Pull Through Cache**: Cache repositories from upstream registries with periodic synchronization to ensure up-to-date images
- **Managed Signing**: Automatic cryptographic signature generation for pushed images using AWS Signer

- **Repository Creation Templates**: Define settings for automatically created repositories including encryption, policies, and tags
- **Encryption**: Image encryption at rest using AWS KMS and in-transit encryption for secure image storage
- **IAM Integration**: Fine-grained access control using AWS IAM roles and policies for repository and image access

Value Proposition

Amazon ECR eliminates the need to operate and manage container registries or underlying infrastructure, providing a fully managed solution with deep AWS integration. ECR offers enterprise-grade security with encryption at rest and in-transit, vulnerability scanning, and IAM-based access controls:

- The service provides high availability across multiple availability zones with automatic scaling and 99.9% uptime SLA
- ECR's native integration with Amazon ECS, EKS, and Fargate streamlines container workflows, while cross-region replication and pull-through caching optimize global deployments
- Cost-effective pricing with no upfront fees, pay-as-you-use model, and AWS Free Tier inclusion make it economical for organizations of all sizes

Service Integration

Amazon ECR integrates natively with core AWS container orchestration services including Amazon Elastic Container Service (ECS), Amazon Elastic Kubernetes Service (EKS), and AWS Fargate for seamless container deployment workflows. ECR works with AWS Lambda for serverless container functions and integrates with AWS CodePipeline, CodeBuild, and CodeDeploy for CI/CD automation:

- The service leverages AWS IAM for access management, AWS KMS for encryption, Amazon CloudWatch for monitoring and logging, and AWS CloudTrail for audit trails
- ECR also integrates with third-party tools like Jenkins, GitHub Actions, GitLab, and other popular CI/CD platforms through standard Docker Registry API compatibility

Onboarding and Offboarding

Getting started with Amazon ECR requires an AWS account and the latest AWS CLI version installed. Users begin by creating a repository using the AWS Management Console, CLI, or SDKs, then authenticate Docker CLI to ECR using the get-authorization-token command:

- The basic workflow involves building Docker images locally, tagging them with the ECR repository URI, and pushing images to ECR using standard Docker commands
- ECR provides comprehensive getting started guides with step-by-step tutorials for creating repositories, pushing first images, and integrating with ECS or EKS

- The service offers both console-based and programmatic access methods to accommodate different user preferences and automation requirements

Getting Started Guide:
https://docs.aws.amazon.com/AmazonECR/latest/userguide/getting-started-cli.html

Data Removal

Offboarding from Amazon ECR involves deleting container images and repositories to remove all stored data. Users can delete individual images using the console, CLI, or SDKs, which removes both the image and associated tags. Repository deletion requires the –force flag if images exist, permanently removing all contained images and metadata. Lifecycle policies can automate cleanup before offboarding. Cross-region replicated images must be deleted from each region separately. Once repositories are deleted, all associated data including images, tags, scan results, and policies are permanently removed from AWS infrastructure.

Backup/Restore and Disaster Recovery

Amazon ECR provides built-in backup and disaster recovery through cross-region replication, automatically copying images to secondary regions for redundancy. Images stored in ECR are replicated across multiple availability zones within a region for high availability. ECR supports native multi-region replication configured at the registry level, enabling automated disaster recovery scenarios. The service integrates with AWS Application Recovery Controller for automated failover orchestration. ECR's durability is enhanced by AWS S3-backed storage with 99.999999999% (11 9's) durability, providing enterprise-grade data protection without additional backup infrastructure.

Backup Capabilities

Amazon ECR includes built-in backup capabilities through cross-region replication that automatically copies container images to designated regions. While ECR doesn't directly integrate with AWS Backup service, its native replication provides automated backup functionality:

- Images are stored with high durability using AWS S3 infrastructure, ensuring data persistence
- Organizations can implement additional backup strategies using ECR APIs to programmatically copy images to multiple regions or accounts
- ECR's lifecycle policies can be configured to maintain image versions while cleaning up older copies for cost optimization

Disaster Recovery

Amazon ECR supports disaster recovery through cross-region and cross-account replication capabilities that can be configured at the registry level. While ECR doesn't

directly integrate with AWS Elastic Disaster Recovery service, it provides robust disaster recovery through automated image replication and high availability architecture:

- ECR repositories can be replicated to multiple regions simultaneously, enabling rapid failover scenarios
- The service works with AWS Application Recovery Controller to orchestrate complete application disaster recovery including container image availability
- ECR's multi-AZ architecture within regions provides automatic failover for high availability

Recovery Objectives

Amazon ECR helps customers achieve aggressive RPO and RTO objectives through real-time cross-region replication that continuously synchronizes images with near-zero data loss. ECR's multi-AZ architecture provides RTO of minutes within a region through automatic failover. Cross-region replication enables RPO of seconds to minutes depending on image size and network conditions. Integration with AWS Application Recovery Controller enables automated disaster recovery orchestration with RTO objectives measured in minutes. ECR's high availability design and global replication capabilities support enterprise-grade recovery requirements for containerized applications.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/AmazonECR/latest/userguide/service-quotas.html

FAQ: https://aws.amazon.com/ecr/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/AmazonECR/latest/userguide/what-is-ecr.html

User Guide: https://docs.aws.amazon.com/AmazonECR/latest/userguide/

API Reference: https://docs.aws.amazon.com/AmazonECR/latest/APIReference/

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/ecr/pricing/

Cost Optimization

Amazon ECR offers several unique cost optimization strategies including lifecycle policies that automatically delete unused or old images based on configurable rules for age, count,

or tags. The service provides separate pricing for private and public repositories, with public repositories offering significant bandwidth allowances:

- ECR's pay-as-you-use model charges only for storage used and data transferred, with no upfront costs or minimum commitments
- Cross-region replication can be optimized by replicating only necessary images to reduce storage and transfer costs
- The AWS Free Tier includes 500 MB monthly storage for private repositories and 50 GB for public repositories, enabling cost-effective experimentation and small-scale deployments

Savings Considerations

Key cost savings considerations for Amazon ECR include implementing lifecycle policies to automatically remove unused images, preventing storage cost accumulation over time. Using public repositories for open-source or widely distributed images eliminates storage charges while leveraging ECR's global content delivery:

- Organizations can optimize data transfer costs by strategically placing repositories in regions closest to container workloads and using VPC endpoints to eliminate internet gateway charges
- Consolidating repositories and using image layer caching reduces storage through deduplication
- Regular auditing using ECR usage reports helps identify cost optimization opportunities, while leveraging AWS Free Tier allowances and volume discounts for larger deployments can significantly reduce overall container registry costs

Service Name

Amazon Elastic Container Service (ECS)

Service Overview

Amazon Elastic Container Service (Amazon ECS) is a fully managed container orchestration service that helps you easily deploy, manage, and scale containerized applications. As a fully managed service, Amazon ECS comes with AWS configuration and operational best practices built-in. It's integrated with both AWS tools, such as Amazon Elastic Container Registry, and third-party tools, such as Docker. This integration makes it easier for teams to focus on building the applications, not the environment. You can run and scale your container workloads across AWS Regions in the cloud, and on-premises, without the complexity of managing a control plane.

Key Use Cases

- Web applications and APIs: Long-running stateless applications that require high availability and automatic scaling
- Batch processing: Running containerized batch jobs for data processing, ETL operations, and computational workloads
- Microservices architecture: Breaking down applications into smaller, independently deployable services
- Hybrid deployments: Running containerized workloads across AWS and on-premises infrastructure using ECS Anywhere
- Modernizing legacy applications: Containerizing existing applications to improve portability and scalability
- Machine learning workloads: Running ML training and inference workloads with GPU acceleration support

Features

- **Task Definitions**: Blueprint for applications that specifies container configurations, resource requirements, and networking
- **Clusters**: Logical grouping of compute resources where containers run
- **Services**: Long-running stateless applications with automatic scaling and health monitoring
- **Tasks**: Single instances of applications such as batch jobs that perform work and then stop
- **Auto Scaling**: Automatic scaling of both container instances and services based on demand
- **Service Connect**: Service discovery and connectivity with traffic monitoring for inter-service communication

- **Blue/Green Deployments**: Built-in deployment strategy for safer software updates with zero downtime
- **ECS Anywhere**: Extends ECS to run containers on on-premises infrastructure
- **Express Mode**: Simplified deployment that automatically provisions infrastructure components
- **Multiple Launch Types**: Support for Fargate serverless, EC2 instances, and ECS Managed Instances

Value Proposition

Amazon ECS provides a fully managed container orchestration experience that eliminates infrastructure management overhead while maintaining flexibility and control. Unlike self-managed Kubernetes, ECS integrates seamlessly with AWS services and provides built-in security, monitoring, and scaling capabilities:

- It offers multiple compute options from serverless Fargate to EC2 instances, allowing customers to optimize for cost and performance
- ECS reduces operational complexity through features like Express Mode for rapid deployment, automated infrastructure optimization with Managed Instances, and native blue/green deployments for safe software updates

Service Integration

Amazon ECS integrates deeply with core AWS services to provide a comprehensive container platform. It works with Amazon ECR for container image storage, Elastic Load Balancing for traffic distribution, and AWS Cloud Map for service discovery:

- VPC integration provides secure networking, while IAM enables fine-grained access control
- CloudWatch offers monitoring and logging capabilities, and AWS Systems Manager supports configuration management
- ECS also integrates with AWS Fargate for serverless compute, Amazon EC2 for traditional instances, and supports hybrid deployments through integration with AWS Outposts and on-premises infrastructure via ECS Anywhere

Onboarding and Offboarding

Customers can get started with Amazon ECS through multiple approaches. The Express Mode provides the fastest path by automatically provisioning all required infrastructure with a single command:

- Traditional setup involves creating clusters, defining task definitions, and configuring services through the AWS Management Console, CLI, or Infrastructure as Code tools like CloudFormation and CDK

- AWS provides comprehensive tutorials, including hands-on labs with AWS Copilot for simplified container deployment
- The service supports gradual adoption, allowing teams to start with simple applications and expand to complex microservices architectures as their expertise grows

Getting Started Guide:
https://docs.aws.amazon.com/AmazonECS/latest/developerguide/Welcome.html

Data Removal

To offboard from Amazon ECS, customers delete services, task definitions, and clusters through the console or API. When deleting services, ECS automatically scales them to zero tasks before removal. Container instances are deregistered, and external instances (ECS Anywhere) require cleanup of AWS resources including stopping ECS and Systems Manager services. Customers should remove ECR repositories, load balancers, and related CloudFormation stacks. ECS provides automated cleanup features for stopped tasks and unused Docker images on container instances.

Backup/Restore and Disaster Recovery

Amazon ECS supports disaster recovery through multi-AZ deployments and cross-region replication strategies. Container images stored in Amazon ECR can be replicated across regions. Task definitions and service configurations can be backed up using Infrastructure as Code. For stateful applications, customers should implement backup strategies for persistent data using EBS snapshots or database replication. ECS integrates with AWS Backup for centralized backup management of associated resources.

Backup Capabilities

Amazon ECS does not have native backup capabilities as it primarily manages stateless containers. However, it integrates with AWS Backup for backing up associated resources like EBS volumes attached to container instances:

- Container images are stored in Amazon ECR with cross-region replication support
- Service configurations and task definitions can be version-controlled and backed up through Infrastructure as Code tools like CloudFormation and CDK

Disaster Recovery

Amazon ECS supports disaster recovery through architectural patterns rather than direct integration with AWS Elastic Disaster Recovery. Customers implement DR using multi-region deployments, where ECS services can be replicated across regions:

- ECS works with Application Load Balancers and Route 53 for traffic failover
- The service's stateless nature facilitates quick recovery by launching new tasks in healthy regions when failures occur

Recovery Objectives

Amazon ECS helps achieve low RPO and RTO objectives through its distributed architecture and integration with AWS services. Multi-AZ deployments provide sub-minute RTO for availability zone failures. Cross-region deployments can achieve RTOs of minutes with proper automation. RPO depends on data replication strategies for persistent storage, with ECS supporting real-time replication through database services and storage solutions. Auto Scaling and health checks enable rapid detection and recovery from failures.

Service Constraints

Service Quotas:
https://docs.aws.amazon.com/AmazonECS/latest/developerguide/service-quotas.html

FAQ: https://aws.amazon.com/ecs/faqs/

Technical Requirements

Service Documentation:
https://docs.aws.amazon.com/AmazonECS/latest/developerguide/

User Guide:
https://docs.aws.amazon.com/AmazonECS/latest/developerguide/Welcome.html

API Reference: https://docs.aws.amazon.com/AmazonECS/latest/APIReference/

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/ecs/pricing/

Cost Optimization

Amazon ECS offers several unique cost optimization features including AWS Fargate Spot for up to 70% savings on interruptible workloads, ECS Managed Instances that automatically optimize infrastructure by right-sizing and consolidating tasks, and intelligent capacity management that removes idle instances. Compute Optimizer provides recommendations for ECS services running on Fargate:

- Customers can mix Fargate and EC2 launch types within the same cluster, use Savings Plans for predictable workloads, and leverage Auto Scaling to match capacity with demand
- Express Mode automatically shares Application Load Balancers across multiple services to reduce costs

Savings Considerations

Primary cost savings come from choosing the right compute option for workloads: Fargate for variable workloads eliminates unused capacity costs, while EC2 instances with Reserved Instances or Savings Plans provide savings for steady-state workloads. ECS Managed Instances optimize costs through automatic instance right-sizing and task consolidation:

- Using Fargate Spot can reduce costs by up to 70% for fault-tolerant applications
- Proper resource allocation through CPU and memory optimization prevents over-provisioning
- Auto Scaling prevents paying for unused capacity during low-demand periods, while shared load balancers in Express Mode distribute infrastructure costs across multiple services

Service Name

Amazon Elastic File System (EFS)

Service Overview

Amazon Elastic File System (EFS) is a serverless, fully elastic file storage service that provides scalable, shared file storage for AWS compute services. It automatically scales from gigabytes to petabytes without disrupting applications and supports the NFSv4.1 and NFSv4.0 protocols. EFS eliminates the need to provision or manage storage capacity and performance, offering a simple web services interface for creating and configuring file systems. The service manages all file storage infrastructure, including deployment, patching, and maintenance of complex configurations, making it accessible across EC2, ECS, EKS, Lambda, and Fargate.

Key Use Cases

- Big Data and Analytics: Processing large datasets with high throughput and concurrent access from multiple compute instances
- Content Management and Web Serving: Sharing website content, media files, and documents across web servers with low latency access
- Media Processing Workflows: Supporting video encoding, image processing, and other media operations requiring shared file access
- Application Development and Testing: Providing shared storage for development environments, CI/CD pipelines, and testing frameworks
- Database Backups and Disaster Recovery: Storing database backups and supporting cross-region replication for business continuity
- Home Directories and User Data: Serving as centralized storage for user files and application data across multiple instances

Features

- **Elastic Scaling**: Automatically scales from gigabytes to petabytes without disrupting applications
- **Multiple Storage Classes**: Standard, Infrequent Access (IA), and Archive classes with lifecycle management
- **Regional and One Zone Types**: Regional for high durability across AZs, One Zone for cost-optimized single AZ storage
- **Performance Modes**: General Purpose and Max I/O modes for different latency and throughput requirements
- **Throughput Modes**: Elastic, Provisioned, and Bursting throughput modes for workload optimization
- **Encryption**: Data encryption at rest and in transit using AWS KMS

- **EFS Replication**: Cross-region replication for disaster recovery with RPO of 15 minutes
- **Access Points**: Secure, fine-grained access control to specific directories and files
- **AWS Backup Integration**: Automated backup and restore capabilities with configurable retention policies

Value Proposition

Amazon EFS offers significant advantages over alternatives through its serverless, fully managed approach that eliminates infrastructure management overhead. Unlike self-managed solutions, EFS provides 99.999999999% durability with automatic multi-AZ replication, delivering up to 219% ROI according to Forrester research:

- The service offers elastic scaling without capacity planning, pay-as-you-go pricing with no upfront costs, and seamless integration with AWS compute services
- EFS provides cost optimization through intelligent tiering, reducing storage costs by up to 94% with Infrequent Access classes, while maintaining high performance with sub-millisecond latency and support for thousands of concurrent connections

Service Integration

EFS integrates deeply with core AWS compute services including Amazon EC2 for persistent shared storage, AWS Lambda for serverless file access, Amazon ECS and EKS for containerized workloads, and AWS Fargate for serverless containers. It works with AWS Backup for automated data protection, AWS DataSync for data transfer, and AWS Transfer Family for managed file transfers:

- EFS connects with CloudWatch for monitoring, CloudFormation for infrastructure as code, CloudTrail for API logging, and IAM for access control
- The service supports on-premises connectivity through AWS Direct Connect and VPN, enabling hybrid cloud architectures and data migration scenarios

Onboarding and Offboarding

Customers can get started with EFS through multiple approaches: using the EC2 launch wizard for one-click file system creation during instance launch, the EFS console for custom configurations, or AWS CLI/API for programmatic deployment. The Quick Create feature automatically configures recommended settings including Regional file system type, General Purpose performance mode, and Elastic throughput:

- Prerequisites include an AWS account, VPC setup, and appropriate security groups
- The process involves creating the file system, configuring mount targets in desired subnets, and mounting via NFS protocol
- AWS provides comprehensive documentation, tutorials, and getting started guides to facilitate onboarding

Getting Started Guide: https://docs.aws.amazon.com/efs/latest/ug/getting-started.html

Data Removal

Offboarding from EFS involves deleting file systems and associated resources through the EFS console, AWS CLI, or API. Customers must first unmount file systems from all instances, then delete mount targets in each Availability Zone. File system deletion is irreversible and permanently removes all data, so AWS recommends creating backups before deletion. Access points, replication configurations, and lifecycle policies should be removed separately. AWS Backup retains backups according to configured retention policies even after file system deletion.

Backup/Restore and Disaster Recovery

EFS provides comprehensive backup and disaster recovery through AWS Backup integration with automatic daily backups enabled by default for console-created file systems. EFS Replication offers cross-region disaster recovery with continuous synchronization and 15-minute RPO for most file systems. Backup rates reach 2,000 files per second or 400 MBps, while restore rates achieve 1,500 files per second or 200 MBps. Point-in-time recovery and incremental backups minimize storage costs and recovery times.

Backup Capabilities

EFS fully integrates with AWS Backup for managed backup services, performing incremental backups that retain necessary reference data for full restoration. Customers can configure automatic backup policies or create on-demand backups without backup plans:

- Regional file systems automatically replicate data across multiple Availability Zones, while One Zone file systems include automatic daily backups by default
- Backup duration can extend up to 30 days without consuming burst credits or affecting file operations

Disaster Recovery

EFS supports disaster recovery through EFS Replication for cross-region protection and works with AWS Elastic Disaster Recovery for comprehensive DR strategies. EFS Replication provides continuous data synchronization to destination regions with configurable RPO and RTO objectives. The service enables failover scenarios by making replica file systems writeable during disasters, with options for failback to original regions while preserving or discarding changes made during recovery periods.

Recovery Objectives

EFS helps customers achieve aggressive RPO and RTO objectives through multiple mechanisms: EFS Replication maintains 15-minute RPO for cross-region disaster recovery,

while Regional file systems provide continuous availability even during single AZ outages. Automated backups support point-in-time recovery scenarios, and the service's high availability architecture minimizes downtime. Fast restore capabilities and elastic scaling ensure rapid recovery from backup or replica systems during disaster scenarios.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/efs/latest/ug/limits.html

FAQ: https://aws.amazon.com/efs/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/efs/latest/ug/whatisefs.html

User Guide: https://docs.aws.amazon.com/efs/latest/ug/index.html

API Reference: https://docs.aws.amazon.com/efs/latest/ug/api-reference.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/efs/pricing/

Cost Optimization

EFS offers several unique cost optimization features including three storage classes (Standard, Infrequent Access, Archive) with automatic lifecycle management that can reduce costs by up to 94%. EFS Intelligent-Tiering automatically moves files between storage classes based on access patterns without performance impact:

- One Zone file systems provide cost savings for workloads not requiring multi-AZ resilience
- Elastic throughput mode eliminates over-provisioning by automatically scaling performance, while storage optimization through deduplication and compression reduces storage footprint
- Customers can share file systems across multiple applications to maximize throughput utilization

Savings Considerations

Primary cost savings opportunities include leveraging Infrequent Access and Archive storage classes for rarely accessed data, implementing One Zone file systems for development/testing environments, and using lifecycle policies to automatically transition files to lower-cost tiers. Sharing EFS file systems across multiple applications maximizes resource utilization without additional costs:

- Elastic throughput eliminates provisioned throughput charges during low-usage periods, while proper mount target placement minimizes cross-AZ data transfer charges
- Regular monitoring and rightsizing of performance modes ensures customers pay only for required performance levels rather than over-provisioning resources

Service Name

Amazon Elastic Kubernetes Service (EKS)

Service Overview

Amazon Elastic Kubernetes Service (EKS) is a fully managed Kubernetes service that simplifies running Kubernetes on AWS without the need to install, operate, or maintain your own Kubernetes control plane or nodes. EKS automatically manages the availability and scalability of the Kubernetes control plane nodes responsible for scheduling containers, managing application availability, storing cluster data, and other key tasks. The service provisions and scales the control plane across multiple AWS Availability Zones for high availability and fault tolerance, automatically detecting and replacing unhealthy control plane nodes while providing automatic version updates and patching.

Key Use Cases

- Microservices architecture: Building and deploying cloud-native applications using microservices patterns with Kubernetes service discovery
- Machine learning workloads: Running ML training and inference workloads with popular frameworks and GPU support on scalable infrastructure
- Batch processing: Executing cost-effective batch processing and big data workloads using EC2 Spot Instances for compute optimization
- Hybrid deployments: Deploying applications across cloud and on-premises environments using EKS Hybrid Nodes and AWS Outposts
- High-availability applications: Building resilient applications with Elastic Load Balancing integration and multi-AZ deployment capabilities
- Platform engineering: Creating custom Kubernetes APIs and abstractions for standardized application deployment patterns

Features

- **Managed Control Plane**: Fully managed Kubernetes control plane with automatic scaling, patching, and high availability across multiple AZs
- **Amazon EKS Auto Mode**: Serverless compute for containers that automatically manages both control plane and worker nodes with automated patching and scaling
- **EKS Capabilities**: Fully managed cluster features including Argo CD for GitOps, AWS Controllers for Kubernetes (ACK), and Kube Resource Orchestrator (KRO)
- **Multiple Compute Options**: Support for Amazon EC2, AWS Fargate, Nitro and Graviton instances, EC2 Spot Instances, and managed node groups
- **Advanced Networking**: VPC native networking, IPv6 support, load balancing integration, and application networking capabilities

- **Security and Compliance**: EKS Pod Identity, IAM for RBAC, encryption at rest and in transit, compliance certifications, and container image signature verification
- **Hybrid Deployment Support**: Amazon EKS Anywhere, EKS Hybrid Nodes, AWS Outposts integration, and EKS Connector for external cluster management
- **Observability and Monitoring**: Integration with Amazon Managed Service for Prometheus, CloudWatch Container Insights, logging, and cost allocation tagging

Value Proposition

Amazon EKS provides significant operational advantages over self-managed Kubernetes by eliminating the complexity of control plane management while maintaining full Kubernetes compatibility. Organizations benefit from reduced operational overhead as AWS handles control plane provisioning, scaling, patching, and high availability across multiple Availability Zones:

- The service offers enterprise-grade security with deep AWS service integrations for IAM, VPC, and compliance frameworks
- EKS supports diverse deployment patterns from cloud-native to hybrid environments with EKS Anywhere and Hybrid Nodes
- The fully managed approach allows teams to focus on application development rather than infrastructure management while maintaining the flexibility to use existing Kubernetes tools and applications

Service Integration

Amazon EKS integrates deeply with core AWS services to provide a comprehensive container platform. Key integrations include Amazon EC2 and AWS Fargate for compute, Amazon EBS and Amazon EFS for persistent storage, and Amazon ECR for container image management:

- Security integrations span AWS IAM for authentication and authorization, Amazon VPC for network isolation, and AWS CloudTrail for audit logging
- Monitoring and observability are enhanced through Amazon CloudWatch, Amazon Managed Service for Prometheus, and AWS X-Ray
- Load balancing integrates with Elastic Load Balancing (ALB, NLB) while Amazon GuardDuty provides threat detection
- Additional integrations include AWS Controllers for Kubernetes (ACK) for managing AWS resources, AWS Backup for cluster protection, and AWS App Mesh for service mesh capabilities

Onboarding and Offboarding

Customers can get started with Amazon EKS using multiple approaches. The quickest path involves using eksctl, AWS's command-line utility that creates clusters with a single command, automatically configuring necessary IAM roles and VPC settings:

- Alternatively, customers can use the AWS Management Console, AWS CLI, or infrastructure-as-code tools like CloudFormation and Terraform
- EKS Auto Mode simplifies the experience further by automatically managing both control plane and worker nodes
- Prerequisites include setting up AWS CLI, kubectl, and appropriate IAM permissions
- AWS provides comprehensive getting started guides, workshops, and sample applications to accelerate adoption
- The service supports both Fargate for serverless containers and EC2-based managed node groups for traditional deployments

Getting Started Guide: https://docs.aws.amazon.com/eks/latest/userguide/getting-started.html

Data Removal

To offboard from Amazon EKS, customers must first delete all workloads and persistent volumes to ensure data removal, then delete node groups and the EKS cluster itself. The process involves removing Kubernetes resources, deleting associated AWS resources like Load Balancers and EBS volumes, and finally terminating the cluster through the AWS console, CLI, or eksctl. Data in persistent volumes attached via EBS, EFS, or S3 must be explicitly deleted as these resources persist beyond cluster deletion. AWS provides detailed cleanup procedures to prevent orphaned resources and unexpected charges.

Backup/Restore and Disaster Recovery

Amazon EKS integrates with AWS Backup to provide comprehensive cluster protection including both Kubernetes cluster state and persistent storage. Backups capture Kubernetes manifests, secrets, config maps, and persistent volumes from EBS, EFS, and S3. The service creates composite recovery points enabling full cluster restoration or selective namespace recovery. EKS supports automated backup scheduling, cross-region and cross-account copies, and immutable backup vaults for compliance requirements.

Backup Capabilities

EKS provides native integration with AWS Backup for automated, centralized backup management. The service supports both on-demand and scheduled backups of cluster state and persistent data without requiring agents or add-ons:

- Backup capabilities include incremental backups, lifecycle management, encryption, and cross-region replication
- AWS Backup creates composite recovery points allowing granular restoration options from full clusters to individual persistent volumes

Disaster Recovery

Amazon EKS supports disaster recovery through multiple mechanisms including AWS Backup integration, multi-region deployment capabilities, and AWS Elastic Disaster Recovery integration. The service enables pilot light, warm standby, and active-active disaster recovery strategies. EKS clusters can be restored across regions and accounts, while persistent data protection spans EBS snapshots, EFS replication, and S3 cross-region replication for comprehensive DR coverage.

Recovery Objectives

EKS helps customers meet RPO and RTO objectives through automated backup scheduling aligned with recovery point requirements, cross-region cluster and data replication for faster recovery, and multi-AZ control plane architecture for high availability. The service supports various DR strategies from backup/restore (higher RPO/RTO, lower cost) to active-active deployments (minimal RPO/RTO, higher cost). Integration with AWS Backup enables policy-driven protection with configurable retention periods.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/eks/latest/userguide/service-quotas.html

FAQ: https://aws.amazon.com/eks/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/eks/latest/userguide/index.html

User Guide: https://docs.aws.amazon.com/eks/latest/userguide/index.html

API Reference: https://docs.aws.amazon.com/eks/latest/APIReference/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/eks/pricing/

Cost Optimization

Amazon EKS offers multiple cost optimization strategies including EKS Auto Mode for serverless compute with automatic rightsizing, EC2 Spot Instances integration for batch workloads with up to 90% cost savings, and Graviton-based instances for better price-performance. The service supports Compute Savings Plans and Reserved Instances for predictable workloads:

- Cost allocation tagging and integration with AWS Cost Explorer and third-party tools like Kubecost enable granular cost tracking and optimization
- Cluster Autoscaler and Vertical Pod Autoscaler help optimize resource utilization, while topology-aware routing reduces cross-AZ data transfer costs

Savings Considerations

Key cost savings opportunities include using mixed instance types and Spot Instances for non-critical workloads, implementing proper resource requests and limits to avoid over-provisioning, and leveraging EKS Auto Mode to eliminate idle node costs. Customers should optimize networking costs through VPC endpoints and topology-aware routing, use ECR image replication for reduced data transfer, and implement cluster consolidation strategies. Cost monitoring through Kubecost integration and AWS native tools helps identify waste, while appropriate node group sizing and utilizing cheaper storage classes for persistent volumes can significantly reduce operational expenses.

Service Name

Amazon Elastic MapReduce (EMR)

Service Overview

Amazon EMR is a managed big data platform that simplifies running big data frameworks such as Apache Hadoop, Apache Spark, Apache Hive, Apache Flink, Trino, and Presto on AWS. It automates the provisioning, configuration, and scaling of clusters, allowing organizations to process and analyze vast amounts of data cost-effectively. EMR handles the complex setup of distributed computing environments while providing integration with AWS services like S3, DynamoDB, and Kinesis for comprehensive data processing workflows.

Key Use Cases

- Log processing and clickstream analysis: Processing large volumes of web logs and user behavior data for analytics and insights
- Extract, Transform, Load (ETL) operations: Moving and transforming data between various AWS data stores and preparing data for analytics
- Machine learning and predictive analytics: Running ML algorithms at scale using frameworks like Spark MLlib for model training and inference
- Interactive data analysis and business intelligence: Performing ad-hoc queries and generating reports using SQL engines like Presto and Spark SQL
- Genomics and scientific computing: Processing large-scale genomic datasets and running complex scientific simulations
- Real-time stream processing: Analyzing streaming data from Kinesis and other sources for real-time insights and decision making

Features

- **Managed Scaling**: Automatically scales clusters up and down based on workload demands with configurable policies for optimal performance and cost efficiency
- **Multiple Deployment Options**: Offers EMR on EC2, EMR Serverless, EMR on EKS, and EMR on Outposts to meet different operational requirements
- **EMR Studio**: Web-based integrated development environment for data scientists and engineers with Jupyter notebook support and collaborative features
- **Open-Source Framework Support**: Pre-configured support for Apache Spark, Hadoop, Hive, Flink, Presto, HBase, and many other popular big data frameworks
- **Instance Fleet Management**: Supports mixing On-Demand and Spot instances across multiple instance types and Availability Zones for cost optimization and resilience

- **Security Integration**: Built-in integration with IAM, VPC, encryption at rest and in transit, and AWS Lake Formation for fine-grained access control
- **Data Store Flexibility**: Native integration with S3, HDFS, DynamoDB, and other data stores with optimized connectors for efficient data access

Value Proposition

Amazon EMR provides unmatched ease of use by eliminating the complexity of setting up and managing big data infrastructure. Organizations benefit from significant cost savings through per-second billing, Spot instance integration, and automatic scaling capabilities:

- EMR offers superior performance with AWS-optimized distributions of open-source frameworks and native integration with AWS services
- The platform provides enterprise-grade security, reliability, and compliance features while maintaining the flexibility to use preferred open-source tools
- EMR's serverless option further reduces operational overhead, allowing teams to focus on data insights rather than infrastructure management

Service Integration

Amazon EMR deeply integrates with the broader AWS ecosystem for comprehensive data workflows. S3 serves as the primary data lake with optimized connectors (EMRFS) for high-performance data access:

- Integration with AWS Glue provides centralized metadata management and ETL capabilities
- Lake Formation enables fine-grained access control and data governance
- Kinesis integration supports real-time data processing scenarios
- EMR connects with DynamoDB for NoSQL data processing, CloudWatch for monitoring and alerting, IAM for security, and VPC for network isolation
- Additional integrations include AWS Data Pipeline for workflow orchestration, Step Functions for complex workflows, and SageMaker for machine learning model deployment

Onboarding and Offboarding

Getting started with Amazon EMR requires completing AWS account setup prerequisites including IAM user creation and EC2 key pair generation. Customers can launch their first cluster using the EMR console's quick launch option or follow the comprehensive getting started tutorial:

- The process involves creating an S3 bucket for data storage, uploading sample data and scripts, then launching a cluster with desired applications like Spark or Hadoop
- EMR provides templates and guided workflows for common use cases

- Customers can choose between different deployment options (EC2, Serverless, EKS) based on their operational preferences
- EMR Studio offers an integrated development environment for interactive development and testing

Getting Started Guide: https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-gs.html

Data Removal

When offboarding from Amazon EMR, customers must terminate all active clusters through the EMR console or CLI to stop compute charges. Data stored in HDFS is automatically deleted when clusters terminate, while data in S3 persists and must be explicitly deleted. Customers should delete associated S3 buckets, remove CloudWatch log groups, delete security configurations, and clean up any EMR-created IAM roles or policies. For EMR Serverless, applications must be stopped and deleted. Complete data removal requires deleting all associated AWS resources across all regions where EMR was used.

Backup/Restore and Disaster Recovery

Amazon EMR provides backup and disaster recovery through integration with AWS services rather than built-in capabilities. EMRFS enables data storage in S3 with cross-region replication for disaster recovery. EBS snapshots can backup persistent data on core nodes. EMR clusters can be automatically recreated using infrastructure as code templates. For disaster recovery, clusters can be launched in alternate regions with data accessed from replicated S3 buckets. EMR's stateless compute architecture allows for rapid cluster recreation and recovery.

Backup Capabilities

Amazon EMR does not include native backup capabilities but integrates with AWS Backup for EBS volumes attached to cluster nodes. Primary data protection relies on storing data in Amazon S3, which provides 99.999999999% durability and supports versioning and cross-region replication:

- For applications requiring persistent storage, customers can use EBS snapshots or configure automated backup policies
- EMR's integration with S3 as the primary data store provides inherent data protection and backup capabilities through S3's durability and replication features

Disaster Recovery

Amazon EMR supports disaster recovery strategies through AWS infrastructure rather than integration with AWS Elastic Disaster Recovery service. EMR clusters can be quickly recreated in alternate regions using infrastructure templates and configuration scripts:

- Data recovery relies on S3 cross-region replication and multi-AZ storage
- EMR's architecture supports backup and restore, pilot light, and warm standby disaster recovery patterns
- The service's ability to launch clusters rapidly and process data from S3 enables effective disaster recovery implementations with customizable RTO and RPO objectives

Recovery Objectives

Amazon EMR helps customers achieve recovery objectives through rapid cluster provisioning and S3 integration. RPO objectives are met through S3's continuous replication and versioning capabilities, with near-zero data loss potential. RTO objectives can be optimized by pre-staging cluster configurations, using automation scripts for rapid deployment, and maintaining warm standby clusters in alternate regions. EMR's managed scaling and instance fleet capabilities ensure quick capacity acquisition during recovery scenarios, supporting RTO requirements from minutes to hours based on cluster size and configuration complexity.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-service-limits-what-are.html

FAQ: https://aws.amazon.com/emr/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-what-is-emr.html

User Guide: https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-what-is-emr.html

API Reference: https://docs.aws.amazon.com/emr/latest/APIReference/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/emr/pricing/

Cost Optimization

Amazon EMR offers several unique cost optimization features including per-second billing with one-minute minimum charges, eliminating waste from hourly billing models. Spot

instance integration can provide up to 90% cost savings with intelligent instance fleet management across multiple pools:

- Managed scaling automatically adjusts cluster size based on workload demands, preventing over-provisioning
- EMR Serverless eliminates idle time charges by automatically starting and stopping resources
- Reserved Instances and Savings Plans provide predictable cost savings for steady workloads
- S3 Intelligent-Tiering integration optimizes storage costs, and AWS Graviton instances offer additional price-performance benefits

Savings Considerations

Primary cost savings opportunities include leveraging Spot instances for fault-tolerant workloads, which can reduce compute costs by up to 90%. Right-sizing clusters through instance type optimization and managed scaling prevents over-provisioning:

- Using S3 as primary storage instead of HDFS reduces costs and eliminates storage redundancy expenses
- EMR Serverless eliminates charges for idle cluster time and management overhead
- Implementing auto-termination policies prevents forgotten clusters from accruing charges
- Reserved Instances and Savings Plans provide significant savings for predictable workloads
- Cost allocation tagging and detailed monitoring help identify optimization opportunities and control spending across different projects and teams

Service Name

Amazon Elastic VMware Service

Service Overview

Amazon Elastic VMware Service (Amazon EVS) is a fully managed AWS service that enables organizations to run VMware Cloud Foundation (VCF) environments directly within their Amazon Virtual Private Cloud (VPC) on EC2 bare metal instances. The service allows customers to maintain their existing VMware expertise, tools, and processes while benefiting from AWS's scale, reliability, and broad set of services. Amazon EVS provides a ready-to-use VCF environment that includes VMware ESXi hypervisor, vCenter Server, NSX-T network virtualization, and vSAN storage, enabling customers to deploy fully functional VCF environments in hours without extensive re-architecture.

Key Use Cases

- VMware workload migration: Lift-and-shift migration of on-premises VMware-based virtual machines to AWS without modification or refactoring
- Hybrid cloud operations: Extend on-premises VMware environments to AWS while maintaining consistent operational tools and processes
- Disaster recovery: Implement disaster recovery solutions using familiar VMware technologies like vSphere Site Recovery Manager integrated with AWS services

Features

- **VMware Cloud Foundation Integration**: Full VCF 5.2.1 stack including ESXi, vCenter Server, NSX-T, and vSAN
- **Rapid Deployment**: Deploy fully functional VCF environments in hours with guided workflows
- **Self-Management Control**: Full administrative access to optimize virtualization stack and integrate third-party solutions
- **Flexible Storage Options**: Local vSAN storage and external options like Amazon FSx for NetApp ONTAP
- **Hybrid Connectivity**: Direct Connect and VPN connectivity with VMware HCX support for migration
- **AWS Service Integration**: Native integration with over 200 AWS services while maintaining VMware tools

Value Proposition

Amazon EVS provides a unique value proposition by combining AWS's cloud benefits with familiar VMware tools and processes. Customers can achieve rapid cloud migration

without extensive re-architecture, maintaining existing skills and operational procedures while gaining access to AWS's scale, elasticity, and performance:

- The service offers control over architecture, integration of preferred external solutions, flexible licensing options including bring-your-own-license (BYOL), and choice between self-management or AWS Partner management
- Amazon EVS eliminates the complexity of setting up VMware infrastructure on AWS, provides enterprise-grade reliability, and enables customers to leverage AWS's broad service portfolio for modernization while preserving their VMware investment

Service Integration

Amazon EVS natively integrates with core AWS services to provide comprehensive cloud functionality. Key integrations include Amazon VPC for networking isolation and security, AWS Direct Connect and Site-to-Site VPN for hybrid connectivity, Amazon FSx for NetApp ONTAP for external storage, Elastic Load Balancing for application distribution, and Amazon RDS for database services:

- The service also integrates with AWS Secrets Manager for credential management, Amazon CloudWatch for monitoring and metrics collection, AWS Migration Hub for migration tracking, and AWS Backup for data protection
- Additionally, customers can leverage AWS services like Amazon S3, DynamoDB, and SNS to enhance application functionality while maintaining their VMware-based workloads

Onboarding and Offboarding

Customers begin by completing prerequisites including an active AWS account, appropriate IAM permissions, VMware Cloud Foundation license keys, and EC2 capacity reservations for i4i.metal instances. The process involves creating a VPC with adequate CIDR space, configuring DNS and NTP servers, setting up VPC Route Server with BGP peers, and creating network ACLs:

- Through the AWS console, customers deploy Amazon EVS environments using guided workflows, which automatically provision VCF management appliances and ESXi hosts
- After deployment, customers retrieve VCF credentials from AWS Secrets Manager to access vCenter Server, NSX, and SDDC Manager
- Optional steps include configuring on-premises connectivity via HCX and establishing hybrid network connections through Direct Connect or VPN

Getting Started Guide: https://docs.aws.amazon.com/evs/latest/userguide/getting-started.html

Data Removal

The offboarding process involves deleting all Amazon EVS hosts first, followed by deleting the environment itself through the AWS console or CLI. Customers must ensure all workloads are migrated or backed up before deletion. The service automatically removes VCF management appliances, cleans up associated VPC resources, and deletes stored credentials from AWS Secrets Manager. Additional cleanup includes removing VPC Route Server components, network ACLs, and disassociating subnet route tables to ensure complete resource removal and prevent ongoing charges.

Backup/Restore and Disaster Recovery

Amazon EVS provides backup and disaster recovery capabilities through integration with VMware native technologies and AWS services. The service supports familiar VMware disaster recovery solutions using vSphere Site Recovery Manager and can integrate with third-party backup and disaster recovery solutions. Customers can leverage AWS Backup for protecting AWS resources, implement cross-region replication, and utilize external storage solutions like Amazon FSx for NetApp ONTAP with NetApp BlueXP disaster recovery for comprehensive data protection and recovery scenarios.

Backup Capabilities

Amazon EVS supports backup capabilities through multiple approaches including VMware native backup solutions, third-party integrations, and AWS services. The service allows integration with preferred external backup solutions and supports AWS Backup for protecting associated AWS resources. Customers can implement vSAN-based local backup, utilize Amazon FSx for NetApp ONTAP with native NetApp backup capabilities, and leverage AWS services like Amazon S3 for long-term data retention and cross-region backup strategies.

Disaster Recovery

Amazon EVS supports disaster recovery through VMware-native technologies and AWS integration. The service enables Site Recovery Manager deployments, supports multi-AZ configurations for high availability, and integrates with solutions like NetApp BlueXP disaster recovery. Customers can implement disaster recovery to other AWS regions, utilize AWS Elastic Disaster Recovery for broader protection, and maintain familiar VMware disaster recovery processes while benefiting from AWS's global infrastructure and reliability.

Recovery Objectives

Amazon EVS helps customers meet RPO and RTO objectives through multiple mechanisms including VMware vSphere replication for low RPO requirements, multi-AZ deployments for high availability, and integration with AWS services for automated failover. The service supports Site Recovery Manager for orchestrated disaster recovery, enables

cross-region replication for geographic protection, and provides scalable recovery infrastructure that can rapidly provision resources during disaster scenarios to meet aggressive RTO requirements.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/evs/latest/userguide/service-quotas-evs.html

FAQ: https://aws.amazon.com/evs/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/evs/latest/userguide/index.html

User Guide: https://docs.aws.amazon.com/evs/latest/userguide/index.html

API Reference: https://docs.aws.amazon.com/evs/latest/APIReference/index.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/evs/pricing/

Cost Optimization

Amazon EVS offers several unique cost optimization options including flexible consumption models with on-demand, one-year, and three-year commitment options for predictable cost savings. Customers can bring their own VMware licenses (BYOL) to avoid additional licensing costs, utilize AWS Instance Savings Plans and Reserved Instances for underlying EC2 infrastructure, and optimize storage costs through flexible options including local vSAN and external storage like Amazon FSx. The service supports elastic scaling to match workload demands, enabling customers to add or remove hosts based on capacity requirements and avoid over-provisioning.

Savings Considerations

Main cost savings considerations include leveraging existing VMware license investments through BYOL options, utilizing AWS's pay-as-you-use model without upfront commitments or minimum fees, and avoiding the capital expenditure of on-premises infrastructure. Customers can achieve operational cost savings through reduced datacenter footprint, elimination of hardware refresh cycles, and leveraging AWS's economies of scale. Additional savings come from Migration Acceleration Program (MAP) eligibility, reduced operational overhead through managed infrastructure, and the ability to integrate cost-effective AWS services like S3 for backup and archival storage instead of expensive traditional solutions.

Service Name

Amazon EventBridge

Service Overview

Amazon EventBridge is a serverless event bus service that connects application components through events, enabling scalable event-driven architectures. It provides simple and consistent ways to ingest, filter, transform, and deliver events between applications, AWS services, and third-party software. EventBridge includes event buses for routing events from many sources to many targets, EventBridge Pipes for point-to-point integrations with advanced transformations, and EventBridge Scheduler for serverless task scheduling using cron and rate expressions.

Key Use Cases

- Application Integration: Decoupling microservices and distributed systems by routing events between various application components and services
- IT Automation: Automating operational tasks and workflows by responding to infrastructure changes and system events in real-time
- SaaS Integration: Connecting over 45 SaaS partner applications to AWS services without custom connection code for seamless data flow
- Real-time Stream Processing: Processing and routing streaming data from various sources to multiple targets with content-based filtering

Features

- **Event Buses**: Serverless routers that receive events from multiple sources and deliver them to multiple targets with optional transformations
- **EventBridge Pipes**: Point-to-point integrations for processing events from single source to single target with advanced transformations and enrichment
- **EventBridge Scheduler**: Serverless scheduler for creating, running, and managing tasks using cron and rate expressions with flexible time windows
- **Schema Registry**: Fully managed service for defining, discovering, and sharing event schemas with code generation capabilities
- **API Destinations**: Low-code integrations for sending events to external HTTP endpoints and SaaS applications via REST API calls
- **Event Replay**: Archive and replay past events for debugging, error recovery, and reprocessing scenarios
- **Global Endpoints**: Improve application availability by replicating events across multiple AWS regions for disaster recovery

Value Proposition

EventBridge provides a fully managed, serverless event routing solution that eliminates the need to write custom integration code. It offers superior developer agility through loosely coupled architectures, enhanced end-user experience via real-time event processing, and reduced operational overhead with automatic scaling. Unlike alternatives, EventBridge provides native integration with over 200 AWS services and 45+ SaaS partners, advanced content-based filtering, built-in retry mechanisms, and exactly-once processing guarantees, making it ideal for building reactive, event-driven systems at scale.

Service Integration

EventBridge integrates natively with over 200 AWS services as both event sources and targets. Key integrations include AWS Lambda for serverless compute processing, Amazon SNS and SQS for messaging, Amazon S3 for object storage events, Amazon DynamoDB for database change streams, AWS CloudTrail for API call monitoring, Amazon CloudWatch for system metrics and alarms, AWS Step Functions for workflow orchestration, Amazon Kinesis for data streaming, and AWS Systems Manager for operational tasks. EventBridge also supports cross-account event sharing and integrates with AWS PrivateLink for secure connectivity.

Onboarding and Offboarding

Customers can start using EventBridge immediately through the AWS Management Console, AWS CLI, or SDKs without provisioning infrastructure. Getting started involves creating an event bus, defining rules with event patterns to match incoming events, and specifying targets where matching events should be delivered:

- The service provides tutorials and sample applications to demonstrate common integration patterns
- Customers can leverage the Schema Registry to discover event structures and generate code bindings
- For SaaS integrations, customers simply enable partner event sources and associate them with custom event buses

Getting Started Guide: https://docs.aws.amazon.com/eventbridge/latest/userguide/eb-tutorial-get-started.html

Data Removal

EventBridge offboarding involves deleting event buses, rules, targets, and associated resources through the console, CLI, or APIs. Events are processed in real-time and not persisted by default, so no historical event data removal is required unless using Event Archive feature. Customers should delete archived events, remove IAM roles and policies, disconnect API destinations and connections, and clean up partner event sources.

Schema Registry schemas and discoverers should also be removed to complete the offboarding process.

Backup/Restore and Disaster Recovery

EventBridge provides limited backup capabilities as events are processed in real-time and not persisted by default. The Event Archive feature allows storing events for replay scenarios, serving as a form of backup for historical events. For disaster recovery, EventBridge supports Global Endpoints that automatically replicate events across multiple regions. The service itself is highly available within regions, but customers must implement cross-region replication strategies for comprehensive disaster recovery using multiple event buses and rules across regions.

Backup Capabilities

EventBridge does not have traditional backup capabilities as it processes events in real-time without persistence. However, the Event Archive feature allows capturing and storing events for replay purposes, which serves as a backup mechanism for event history. EventBridge does not integrate directly with AWS Backup service, as it is designed for real-time event processing rather than data storage requiring backup protection.

Disaster Recovery

EventBridge supports disaster recovery through Global Endpoints that automatically failover between regions during outages. The service does not directly integrate with AWS Elastic Disaster Recovery, which focuses on infrastructure replication. Instead, customers implement disaster recovery by deploying EventBridge resources across multiple regions and using Global Endpoints for automatic failover, ensuring continued event processing during regional disruptions.

Recovery Objectives

EventBridge helps achieve low RPO through Global Endpoints that replicate events across regions in near real-time, minimizing data loss during failures. For RTO objectives, the service provides automatic failover capabilities via Global Endpoints and high availability within regions. The Event Replay feature enables recovery from specific points in time when combined with Event Archive, allowing customers to reprocess events from archived history to meet recovery point requirements.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/eventbridge/latest/userguide/eb-quota.html

FAQ: https://aws.amazon.com/eventbridge/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/eventbridge/latest/userguide/

User Guide: https://docs.aws.amazon.com/eventbridge/latest/userguide/

API Reference:
https://docs.aws.amazon.com/eventbridge/latest/APIReference/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/eventbridge/pricing/

Cost Optimization

EventBridge offers several cost optimization strategies including pay-per-use pricing with no upfront costs or minimum fees. The service provides a generous free tier with 5 million custom events monthly:

- Customers can optimize costs by using precise event filtering to reduce unnecessary event processing, leveraging the free AWS service events, and implementing efficient event patterns to minimize rule evaluations
- EventBridge Scheduler offers 14 million free invocations monthly
- Using event transformation at the source rather than targets can reduce data transfer costs

Savings Considerations

Key cost savings come from eliminating custom integration infrastructure and reducing operational overhead through serverless architecture. EventBridge's pay-per-event model scales costs with actual usage, avoiding over-provisioning expenses:

- The service reduces development time and maintenance costs by providing managed integrations with 200+ AWS services and 45+ SaaS partners
- Customers save on networking costs by using EventBridge as a central hub rather than point-to-point integrations
- The free tier and AWS service events provide significant cost advantages for many workloads, especially during development and testing phases

Service Name

Amazon File Cache

Service Overview

Amazon File Cache is a fully managed, high-speed caching service on AWS that enables processing of file data from various data repositories, including on-premises file systems, AWS file systems, and Amazon S3 buckets. It serves as a temporary, high-performance storage location that presents data as a unified set of files and directories. The service provides sub-millisecond latency and high throughput access to dispersed datasets for file-based applications on AWS, supporting hundreds of GBps throughput and millions of IOPS. Amazon File Cache is POSIX-compliant and uses automatic data loading and release mechanisms for optimal performance and cost efficiency.

Key Use Cases

- High performance computing (HPC) workloads: Processing compute-intensive tasks requiring fast access to large datasets distributed across multiple storage locations
- Machine learning and data analytics: Accelerating ML workflows by providing high-speed access to training data stored in S3 or on-premises systems
- Media and content processing: Enabling fast processing of media files and content workflows that require high throughput and low latency access

Features

- **Multi-Repository Support**: Links up to 8 data repositories of the same type (Amazon S3 or NFS) per cache
- **High Performance**: Delivers sub-millisecond latencies, hundreds of GBps throughput, and millions of IOPS
- **POSIX Compliance**: Provides native file system interface with read-after-write consistency and file locking
- **Automatic Data Management**: Lazy-loads data on first access and supports preloading with HSM commands for data export
- **Cache Eviction**: Automatically releases less recently used files when cache fills up, with manual control options
- **Storage Quotas**: Supports user and group quotas for disk space and file count limits
- **Encryption**: Supports encryption at rest and in transit using AWS KMS keys
- **Lustre Integration**: Built on Lustre technology with support for open-source Lustre client and CSI driver

Value Proposition

Amazon File Cache eliminates the complexity of managing high-performance file caching infrastructure while providing exceptional performance with sub-millisecond latencies and massive throughput capabilities. Unlike traditional caching solutions, it offers fully managed operations with automatic scaling, data replication within Availability Zones, and seamless integration with both cloud and on-premises data sources:

- The service provides cost optimization through pay-as-you-use pricing, automatic cache eviction, and eliminates upfront hardware investments
- Its POSIX compliance ensures existing Linux applications work without modifications, while native AWS integrations with services like Batch, EKS, and ECS accelerate deployment of compute-intensive workloads

Service Integration

Amazon File Cache integrates natively with multiple core AWS services to enable comprehensive compute and storage workflows. It connects seamlessly with Amazon S3 buckets and NFS file systems as data repositories, supports access from Amazon EC2 instances across Availability Zones, and works with containerized workloads on Amazon ECS and Amazon EKS through the CSI driver:

- The service integrates with AWS Batch for high-performance computing workloads and AWS Thinkbox Deadline for render farm management
- It leverages AWS KMS for encryption, AWS CloudTrail for API logging, Amazon CloudWatch for performance monitoring, and AWS PrivateLink for secure private network access
- VPC integration provides network isolation and security group controls

Onboarding and Offboarding

Getting started with Amazon File Cache requires an AWS account, Amazon EC2 instance, and appropriate IAM permissions. Customers create a cache through the AWS Console, CLI, or API by specifying storage capacity, subnet IDs, security groups, and data repository associations:

- The Lustre client must be installed on EC2 instances (included with Amazon Linux 2/Amazon Linux, available via AWS repositories for RHEL/CentOS/Ubuntu)
- After mounting the cache using standard Linux commands, applications can access cached data immediately
- The service supports both lazy loading and preloading strategies
- Getting started documentation provides step-by-step guidance including prerequisites, cache creation, client configuration, and running analysis workloads

Getting Started Guide: https://docs.aws.amazon.com/fsx/latest/FileCacheGuide/getting-started.html

Data Removal

Offboarding from Amazon File Cache involves deleting cache resources through AWS Console, CLI, or API using the DeleteFileCache operation. When a cache is deleted, all cached data is permanently lost and cannot be recovered. Customers should export any modified data back to linked data repositories using HSM commands before deletion. The process includes terminating EC2 instances, deleting the cache (which transitions to DELETING status), and optionally removing associated S3 buckets if no longer needed for data preservation.

Backup/Restore and Disaster Recovery

Amazon File Cache does not provide traditional backup capabilities as it serves as a temporary, high-performance cache rather than persistent storage. Data protection relies on the underlying data repositories (S3 or NFS) where the source data resides. The service provides automatic data replication within the same Availability Zone and automatic server replacement if failures occur. For disaster recovery, customers must ensure their linked data repositories have appropriate backup and recovery strategies implemented.

Backup Capabilities

Amazon File Cache does not support traditional backup capabilities and is not integrated with AWS Backup service. As a caching service designed for temporary, high-performance storage, data protection relies on the durability and backup mechanisms of the linked data repositories (Amazon S3 or NFS). The service itself provides automatic replication within the Availability Zone and server replacement for fault tolerance.

Disaster Recovery

Amazon File Cache does not integrate with AWS Elastic Disaster Recovery as it functions as a temporary cache rather than persistent storage. Disaster recovery strategies must focus on the underlying data repositories:

- The service provides automatic data replication within the same Availability Zone and server replacement capabilities
- For multi-region disaster recovery, customers must implement appropriate replication strategies for their S3 buckets or NFS data repositories

Recovery Objectives

Amazon File Cache helps meet RPO and RTO objectives through its automatic replication within Availability Zones and rapid cache recreation capabilities. Since it's a temporary cache with data sourced from repositories, RPO depends on the backup strategy of underlying S3 or NFS data sources. RTO can be minimized through quick cache

provisioning and automatic data loading. However, customers must implement comprehensive disaster recovery planning for their data repositories to achieve specific RPO/RTO targets.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/fsx/latest/FileCacheGuide/limits.html

FAQ: https://aws.amazon.com/filecache/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/fsx/latest/FileCacheGuide/

User Guide: https://docs.aws.amazon.com/fsx/latest/FileCacheGuide/what-is.html

API Reference: https://docs.aws.amazon.com/fsx/latest/APIReference/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/filecache/pricing/

Cost Optimization

Amazon File Cache offers several unique cost optimization features including pay-only-for-provisioned-storage with no upfront costs, automatic cache eviction to free space for new data, and intelligent data loading that reduces unnecessary storage consumption. The service eliminates capital expenditure on caching infrastructure and reduces operational overhead through fully managed operations:

- Customers can optimize costs by right-sizing cache capacity, using lazy loading instead of preloading when appropriate, implementing storage quotas to control usage, and leveraging cache eviction policies
- The temporary nature of cached data reduces long-term storage costs compared to persistent file systems

Savings Considerations

Primary cost savings come from eliminating the need to provision and manage high-performance caching infrastructure, reducing compute resource consumption through faster workload completion times, and optimizing data transfer costs by caching frequently accessed data closer to compute resources. The service reduces operational expenses through automated administration tasks, eliminates hardware refresh cycles, and provides elastic scaling without over-provisioning:

- Pay-as-you-use pricing ensures customers only pay for actual storage consumed, while automatic cache management reduces the need for dedicated storage administration staff
- Integration with spot instances and containerized workloads further enhances cost efficiency for batch processing scenarios

Service Name

Amazon FSx

Service Overview

Amazon FSx is a fully managed file storage service that makes it easy to launch and run popular file systems without hardware provisioning, software configuration, patching, or backups. FSx provides cost-efficient capacity with high levels of reliability and integrates with other AWS services. It offers four file system types: FSx for Lustre (high-performance for HPC/ML workloads), FSx for Windows File Server (Windows-based SMB file shares), FSx for NetApp ONTAP (enterprise features with NFS/SMB support), and FSx for OpenZFS (Linux workloads with ZFS features). Each file system type is optimized for specific workloads and provides enterprise-grade performance, security, and availability features with flexible deployment options including Single-AZ and Multi-AZ configurations.

Key Use Cases

- High Performance Computing (HPC): FSx for Lustre provides up to 1000 GB/s throughput for machine learning, financial modeling, and scientific computing workloads
- Windows workloads: FSx for Windows File Server supports business applications, home directories, SQL Server deployments, SharePoint, and .NET applications with Active Directory integration
- Content management and media processing: All FSx types support content repositories, video processing, and media workflows with high-throughput file access
- Data analytics and machine learning: FSx for Lustre and ONTAP provide high-performance storage for big data analytics, ML training, and data lake architectures
- Enterprise file shares and collaboration: FSx for Windows File Server and ONTAP provide secure, managed file shares with user quotas and access controls
- Backup and archival: Integration with AWS Backup and S3 for automated backup strategies and long-term data retention

Features

- **High Performance Storage**: Sub-millisecond latencies, up to 1000 GB/s throughput for Lustre, hundreds of thousands of IOPS, and scalable performance independent of storage size
- **Multi-Protocol Support**: SMB 2.0-3.1.1, NFS 3.0-4.2, and iSCSI protocols with cross-platform compatibility for Windows, Linux, and macOS clients
- **Data Management**: Data deduplication, compression, intelligent tiering, user/group quotas, snapshots, and automated lifecycle management

- **High Availability**: Multi-AZ deployments with 99.99% SLA, automatic failover, and disaster recovery capabilities across regions
- **Security and Compliance**: Encryption at rest and in transit, Active Directory integration, VPC isolation, file access auditing, and compliance certifications
- **AWS Integration**: Native integration with S3, EC2, ECS, EKS, AWS Backup, DataSync, and other AWS services for hybrid workflows

Value Proposition

Amazon FSx eliminates the complexity and overhead of managing file storage infrastructure while providing enterprise-grade performance and features. Customers benefit from fully managed operations with no hardware provisioning, automatic software updates, and built-in high availability:

- FSx offers significant cost savings through pay-as-you-go pricing, data deduplication (30-80% savings), and intelligent tiering for cold data
- The service provides superior performance with sub-millisecond latencies and massive throughput scaling, while maintaining compatibility with existing applications and protocols
- Deep AWS integration enables seamless hybrid architectures with S3, automated backups, and native support for compute services like EC2 and containers

Service Integration

Amazon FSx integrates natively with core AWS compute services including Amazon EC2, Amazon ECS, and Amazon EKS for file storage access. AWS Backup provides centralized backup management with automated scheduling and retention policies:

- Amazon S3 integration enables data import/export and tiering for cost optimization
- AWS DataSync facilitates migration and synchronization between on-premises and FSx file systems
- AWS Directory Service and Active Directory provide authentication and access control
- VPC networking enables secure access control with security groups and NACLs
- Additional integrations include AWS CloudTrail for auditing, CloudWatch for monitoring, AWS Direct Connect and VPN for hybrid connectivity, and AWS Lambda for automated workflows

Onboarding and Offboarding

Customers begin by setting up prerequisites including an AWS account, VPC configuration, and Active Directory (for Windows File Server). The getting started process involves creating a file system through the AWS Console, CLI, or API by selecting the appropriate file system type, storage capacity, throughput levels, and network configuration:

- After creation, customers mount the file system to compute instances using standard protocols (SMB, NFS) and begin accessing shared storage
- The service provides step-by-step tutorials and documentation for each file system type
- Initial setup typically takes minutes, and customers can immediately begin using enterprise features like snapshots, backups, and user quotas without additional configuration

Getting Started Guide: https://docs.aws.amazon.com/fsx/latest/WindowsGuide/getting-started.html

Data Removal

Customers can delete file systems and all associated data through the AWS Console, CLI, or API. Before deletion, final backups are automatically created and retained according to backup retention policies. All data is permanently removed from FSx storage, and customers are no longer charged for storage or throughput capacity. Existing backups remain available until their retention period expires or they are manually deleted. For complete offboarding, customers should also remove associated resources like mount targets, security groups, and any automated backup policies to ensure no residual costs.

Backup/Restore and Disaster Recovery

Amazon FSx provides comprehensive backup and disaster recovery capabilities including automatic daily backups with configurable retention periods (0-90 days), user-initiated backups retained indefinitely, and integration with AWS Backup for centralized management. All backups are incremental, encrypted, and stored across multiple Availability zones. Cross-region backup copying supports disaster recovery strategies, and point-in-time restore creates new file systems from any backup. FSx also supports native features like Windows shadow copies and NetApp SnapMirror for additional data protection and replication capabilities.

Backup Capabilities

Amazon FSx includes native backup functionality with automatic daily backups enabled by default and user-initiated backups available on-demand. All backups are incremental, file-system-consistent, and encrypted:

- FSx integrates seamlessly with AWS Backup for centralized backup management, automated scheduling, and cross-service backup strategies
- AWS Backup provides additional features like unlimited retention, cross-region backup copying, and compliance reporting
- Both native FSx backups and AWS Backup support point-in-time recovery and restore to new file systems

Disaster Recovery

Amazon FSx supports comprehensive disaster recovery through Multi-AZ deployments with automatic failover, cross-region backup copying, and replication capabilities. While FSx doesn't directly integrate with AWS Elastic Disaster Recovery, customers can implement DR strategies using FSx backup and restore capabilities combined with infrastructure automation. FSx for ONTAP supports NetApp SnapMirror for real-time replication, and customers can use AWS DataSync for scheduled replication between file systems across regions for robust disaster recovery architectures.

Recovery Objectives

Amazon FSx helps customers meet aggressive RPO and RTO objectives through multiple mechanisms. Multi-AZ deployments provide near-zero RTO with automatic failover in minutes and RPO measured in seconds through synchronous replication. Frequent automated backups enable RPO of hours with RTO dependent on restore times. Cross-region backup strategies support longer-term disaster recovery scenarios. FSx for ONTAP's SnapMirror replication can achieve RPO of minutes with RTO based on failover procedures, while intelligent tiering and local caching minimize data loss exposure.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/fsx/latest/WindowsGuide/limits.html

FAQ: https://aws.amazon.com/fsx/windows/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/fsx/

User Guide: https://docs.aws.amazon.com/fsx/latest/WindowsGuide/

API Reference: https://docs.aws.amazon.com/fsx/latest/APIReference/

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/fsx/pricing/

Cost Optimization

Amazon FSx offers several unique cost optimization features including data deduplication that typically saves 30-80% of storage costs by eliminating duplicate data blocks. Intelligent tiering automatically moves infrequently accessed data to lower-cost storage tiers without performance impact:

- HDD storage options provide cost-effective capacity for workloads that don't require SSD performance
- Single-AZ deployments offer approximately 40% cost savings compared to Multi-AZ for non-critical workloads
- User and group quotas prevent storage overuse, while flexible throughput and IOPS provisioning allows customers to pay only for the performance they need
- The service eliminates traditional storage management costs including hardware, maintenance, and administrative overhead

Savings Considerations

Primary cost savings come from eliminating capital expenditure on storage hardware and reducing operational overhead through fully managed operations. Data deduplication provides immediate 30-80% storage cost reduction for typical workloads:

- Intelligent tiering and HDD storage options reduce ongoing costs for cold data
- Pay-as-you-go pricing eliminates overprovisioning waste, while independent scaling of storage and throughput prevents paying for unused performance
- Single-AZ deployments significantly reduce costs for development and non-critical workloads
- Integration with AWS Backup reduces backup infrastructure costs, and efficient data transfer within AWS regions minimizes bandwidth charges
- The elimination of traditional storage maintenance, patching, and management reduces total cost of ownership substantially

Service Name

Amazon FSx for Lustre

Service Overview

Amazon FSx for Lustre is a fully managed, high-performance Lustre file system service designed for applications requiring fast storage and low latency. It provides sub-millisecond latencies, up to multiple terabytes per second of throughput, and up to millions of IOPS. The service eliminates the traditional complexity of setting up and managing Lustre file systems, enabling users to spin up battle-tested high-performance file systems in minutes. FSx for Lustre is POSIX-compliant and works with existing Linux-based applications without modifications, supporting workloads where storage speed matters most.

Key Use Cases

- Machine learning and AI workloads: Training large-scale models requiring high-throughput data processing and low-latency storage access
- High Performance Computing (HPC): Scientific computing, weather forecasting, seismic imaging, genomic analysis, and financial modeling requiring parallel file system performance
- Media and entertainment: Video processing, rendering, and content creation workflows requiring fast file I/O and high bandwidth
- Electronic Design Automation (EDA): Semiconductor design and verification workloads with high metadata operations on numerous small files
- Big data analytics: Processing massive datasets requiring scalable, high-speed file system interface with integration to long-term data repositories

Features

- **Multiple Deployment Options**: Scratch file systems for temporary storage and shorter-term processing, Persistent file systems for longer-term storage with data replication and automatic server replacement
- **Storage Classes**: SSD for high IOPS workloads, HDD for throughput-intensive workloads, Intelligent-Tiering for automatic cost optimization with frequent, infrequent, and archive access tiers
- **Amazon S3 Integration**: Native integration with S3 buckets for seamless data import/export, automatic synchronization, and data repository associations
- **Scalable Metadata Performance**: User-provisioned metadata IOPS up to 15x higher than automatic allocation, independent of storage capacity for workloads with many small files

- **Elastic Fabric Adapter (EFA) Support**: Enhanced network performance for MPI applications and tightly coupled HPC workloads with high-performance networking
- **Data Compression**: LZ4 compression support to reduce storage consumption and optimize costs while maintaining performance

Value Proposition

Amazon FSx for Lustre provides the only fully managed Lustre file system in the cloud, delivering enterprise-grade performance without operational overhead. It offers up to 34% better price-performance than on-premises HDD storage and up to 70% better than other cloud-based Lustre solutions:

- The service eliminates complex setup and management tasks while providing native S3 integration for seamless data workflows
- With multiple deployment options and storage classes, customers can optimize for both performance and cost
- The Intelligent-Tiering storage class provides fully elastic scaling and automatic cost optimization, making it ideal for workloads with varying access patterns

Service Integration

Amazon FSx for Lustre integrates natively with Amazon S3 for data repository access, enabling automatic import/export and synchronization of datasets. It works seamlessly with Amazon EC2 instances, Amazon ECS Docker containers, and Amazon EKS containers for compute access:

- The service integrates with AWS services like Amazon SageMaker AI for machine learning workflows, AWS Batch for batch processing jobs, and AWS ParallelCluster for HPC workloads
- It also supports AWS Backup for centralized backup management, AWS KMS for encryption key management, and Amazon CloudWatch for monitoring and logging
- On-premises integration is available through AWS Direct Connect and VPN connections

Onboarding and Offboarding

Customers can get started by creating an FSx for Lustre file system through the AWS Console, CLI, or API after setting up an AWS account and configuring necessary IAM permissions. The process involves selecting deployment type (scratch or persistent), storage class (SSD, HDD, or Intelligent-Tiering), storage capacity, and throughput requirements:

- Optional configurations include linking to S3 data repositories, provisioning metadata IOPS, and enabling SSD read cache

- Once created, the file system can be mounted on Amazon EC2 instances using standard Lustre client tools
- The service includes comprehensive documentation, getting started guides, and sample applications to accelerate deployment

Getting Started Guide: https://docs.aws.amazon.com/fsx/latest/LustreGuide/getting-started.html

Data Removal

When offboarding, customers can delete FSx for Lustre file systems through the AWS Console, CLI, or API, which permanently removes all data stored on the file system. Before deletion, data can be exported to linked S3 data repositories or backed up using AWS Backup. For persistent file systems, automatic and user-initiated backups provide data protection options. Customers should ensure all mounted clients are unmounted before deletion and consider copying critical data to alternative storage locations.

Backup/Restore and Disaster Recovery

FSx for Lustre supports automatic daily backups and user-initiated backups for persistent file systems not linked to S3 data repositories. Backups are incremental, file-system-consistent, and highly durable with configurable retention periods from 0-90 days. AWS Backup integration provides centralized backup management with cross-region and cross-account backup copying capabilities. Backup restoration creates new file systems with identical configurations, enabling point-in-time recovery for disaster scenarios.

Backup Capabilities

FSx for Lustre provides comprehensive backup capabilities through automatic daily backups during configurable backup windows and on-demand user-initiated backups. Integration with AWS Backup enables centralized backup management, automated scheduling, unlimited retention options, and cross-region/cross-account backup copying. Backups are incremental to minimize storage costs and support restoration to new file systems with identical deployment types and configurations.

Disaster Recovery

Disaster recovery is supported through backup and restore capabilities, cross-region backup copying, and S3 data repository integration. AWS Backup enables cross-region disaster recovery with logically air-gapped vaults for enhanced protection. The service supports restoration from backups to create new file systems in different regions, and S3 integration provides additional data durability and geographic distribution for critical datasets.

Recovery Objectives

FSx for Lustre helps customers achieve RPO objectives through automatic daily backups with configurable retention and real-time S3 synchronization for linked data repositories. RTO objectives are met through rapid backup restoration capabilities, where new file systems can be created from backups with data lazily loaded for faster availability. Cross-region backup copying and S3 integration provide additional recovery options for meeting stringent business continuity requirements.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/fsx/latest/LustreGuide/limits.html

FAQ: https://aws.amazon.com/fsx/lustre/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/fsx/latest/LustreGuide/what-is.html

User Guide: https://docs.aws.amazon.com/fsx/latest/LustreGuide/

API Reference: https://docs.aws.amazon.com/fsx/latest/APIReference/

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/fsx/lustre/pricing/

Cost Optimization

FSx for Lustre offers several unique cost optimization features including Intelligent-Tiering storage class that automatically moves infrequently accessed data to lower-cost tiers, starting at less than $0.005 per GB-month. The service provides scratch file systems for cost-effective temporary storage, HDD storage options for throughput-intensive workloads at lower costs, and data compression with LZ4 to reduce storage consumption:

- Automatic scaling eliminates expensive overprovisioning, and integration with S3 enables cost-effective long-term data archiving
- Customers can optimize metadata IOPS provisioning and choose appropriate throughput tiers based on workload requirements

Savings Considerations

Main cost savings come from eliminating on-premises infrastructure management overhead and capital expenses while providing up to 34% better price-performance than

traditional HDD storage. The pay-as-you-go model with no upfront costs or minimum commitments reduces financial risk:

- Intelligent-Tiering automatically optimizes storage costs based on access patterns, and incremental backups minimize backup storage expenses
- S3 integration enables customers to store active data on high-performance FSx while archiving cold data in cost-effective S3 storage classes
- Scratch file systems provide significant cost savings for temporary workloads compared to persistent storage alternatives

Service Name

Amazon FSx for NetApp ONTAP

Service Overview

Amazon FSx for NetApp ONTAP is a fully managed file storage service that provides high-performance, scalable, and feature-rich shared file storage using NetApp's ONTAP file system. It supports petabyte-scale datasets and up to tens of GBps of throughput per file system. The service offers multi-protocol access through NFS, SMB, iSCSI, and NVMe protocols, automatic data-tiering between SSD and capacity pool storage, and advanced data management features including compression, deduplication, and compaction. It delivers sub-millisecond latencies for SSD storage and provides enterprise-grade capabilities with the agility, scalability, and simplicity of a fully managed AWS service.

Key Use Cases

- Enterprise file shares requiring high performance and multi-protocol access for Linux, Windows, and macOS clients
- Database workloads including Microsoft SQL Server, Oracle, SAP, and other high-performance database applications requiring block storage via iSCSI and NVMe-over-TCP
- Hybrid cloud deployments with on-premises NetApp systems requiring seamless data replication, backup, and disaster recovery using SnapMirror and FlexCache

Features

- **Multi-Protocol Access**: Supports NFS, SMB, iSCSI, and NVMe-over-TCP protocols for broad compatibility
- **Elastic Capacity Pool Tiering**: Automatic data tiering between high-performance SSD and cost-optimized capacity pool storage
- **Storage Efficiency**: Block-level compression, deduplication, and compaction achieving up to 65% storage savings
- **High Availability**: Multi-AZ and Single-AZ deployment options with automatic failover capabilities
- **NetApp Integration**: Native support for SnapMirror replication, FlexCache caching, and NetApp management tools
- **Enterprise Security**: Encryption at rest and in transit, Active Directory integration, antivirus scanning, and access auditing

Value Proposition

Amazon FSx for NetApp ONTAP provides the familiar NetApp ONTAP functionality customers know and trust, delivered as a fully managed AWS service without the

operational overhead of managing on-premises infrastructure. It eliminates the need for upfront capital expenditure on hardware while providing enterprise-grade features like snapshots, clones, and replication:

- The service offers superior cost optimization through elastic capacity pool tiering and storage efficiency features that can reduce storage costs by up to 65%
- Its multi-protocol support ensures compatibility with existing applications across Linux, Windows, and macOS environments
- The seamless integration with existing NetApp environments through SnapMirror and FlexCache enables gradual cloud migration and hybrid architectures without disrupting existing workflows

Service Integration

Amazon FSx for NetApp ONTAP integrates deeply with core AWS compute services including Amazon EC2, Amazon ECS, and Amazon EKS for direct file system mounting. It works seamlessly with AWS Backup for centralized backup management and automated retention policies:

- The service integrates with Amazon CloudWatch for monitoring and alerting, AWS CloudTrail for audit logging, and AWS IAM for access control
- For hybrid scenarios, it connects with AWS Direct Connect and AWS Site-to-Site VPN for secure on-premises connectivity
- It also integrates with Amazon S3 for long-term archival and supports AWS Storage Gateway for hybrid storage workflows
- The service works with AWS DataSync for data migration and Amazon Elastic VMware Service for VMware workloads

Onboarding and Offboarding

Customers start by creating an AWS account and setting up administrative access through AWS IAM Identity Center. The initial file system creation uses the Quick Create option in the Amazon FSx console, which provisions a Multi-AZ or Single-AZ file system with default settings suitable for NFS access from Linux instances:

- For Windows environments, customers can create additional Storage Virtual Machines (SVMs) joined to Microsoft Active Directory for SMB access
- The service provides step-by-step guidance for mounting file systems from Amazon EC2 instances using standard NFS or SMB protocols
- Advanced configurations including throughput capacity, SSD IOPS provisioning, and capacity pool tiering policies can be configured during or after initial creation

Getting Started Guide: https://docs.aws.amazon.com/fsx/latest/ONTAPGuide/getting-started.html

Data Removal

Customers can delete FSx for ONTAP file systems through the AWS console, CLI, or API after ensuring all data is backed up or migrated. The offboarding process involves deleting volumes, storage virtual machines, and finally the file system itself. All automatic backups are retained according to the configured retention period, while user-initiated backups persist until manually deleted. During deletion, customers receive warnings about data loss and must confirm the deletion. The service provides detailed cleanup procedures to remove all associated resources including security groups and endpoints.

Backup/Restore and Disaster Recovery

Amazon FSx for NetApp ONTAP provides comprehensive backup and disaster recovery capabilities through automatic daily backups, user-initiated backups, and native ONTAP features. Automatic backups occur during configurable backup windows with retention periods up to 90 days. The service integrates with AWS Backup for centralized management across multiple AWS services. Native ONTAP features include instantaneous snapshots for point-in-time recovery, SnapMirror for cross-region replication, and FlexCache for distributed caching. All backups are incremental and stored across multiple Availability Zones for durability.

Backup Capabilities

The service includes automatic daily backups enabled by default with configurable retention periods from 0-90 days. User-initiated backups can be created on-demand and are retained indefinitely:

- FSx for ONTAP integrates natively with AWS Backup for centralized backup policies and cross-service management
- The service supports both crash-consistent backups and application-consistent snapshots through native ONTAP snapshot capabilities with up to 1,023 snapshots per volume

Disaster Recovery

Amazon FSx for NetApp ONTAP supports comprehensive disaster recovery through Multi-AZ deployments for high availability within a region and cross-region replication using NetApp SnapMirror technology. The service automatically replicates data across multiple Availability Zones and provides automatic failover capabilities. While it doesn't directly integrate with AWS Elastic Disaster Recovery, it provides native ONTAP disaster recovery features including SnapMirror relationships for automated replication to secondary sites.

Recovery Objectives

FSx for ONTAP helps customers achieve aggressive RPO and RTO objectives through multiple recovery mechanisms. Instantaneous snapshots provide near-zero RPO for point-

in-time recovery, while automatic daily backups ensure RPOs within 24 hours. Multi-AZ deployments provide RTOs in minutes through automatic failover, and cross-region SnapMirror replication enables disaster recovery with RPOs measured in minutes depending on replication frequency. The service's high availability architecture minimizes planned and unplanned downtime.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/fsx/latest/ONTAPGuide/limits.html

FAQ: https://docs.aws.amazon.com/prescriptive-guidance/latest/fsx-ontap-enterprise-deployment/faq.html

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/fsx/latest/ONTAPGuide/what-is-fsx-ontap.html

User Guide: https://docs.aws.amazon.com/fsx/latest/ONTAPGuide/

API Reference: https://docs.aws.amazon.com/fsx/latest/APIReference/

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/fsx/netapp-ontap/pricing/

Cost Optimization

Amazon FSx for NetApp ONTAP offers several unique cost optimization features that set it apart from other storage solutions. The elastic capacity pool tiering automatically moves infrequently accessed data to lower-cost storage while maintaining fast access when needed:

- ONTAP's storage efficiency features including compression, deduplication, and compaction typically achieve 65% storage savings for general-purpose workloads
- The pay-as-you-go pricing model with no upfront costs or licensing fees ensures customers only pay for actual usage
- Thin provisioning allows over-provisioning of volumes without consuming actual storage until data is written
- The service's flexible throughput and IOPS provisioning enables right-sizing for specific performance requirements without over-provisioning

Savings Considerations

Primary cost savings come from eliminating capital expenditure on NetApp hardware and reducing operational overhead through fully managed service delivery. The elastic capacity pool provides significant savings by automatically tiering cold data to cost-optimized storage at a fraction of SSD costs:

- Storage efficiency features can reduce total storage consumption by up to 65% across SSD, capacity pool, and backup storage
- The service eliminates traditional NetApp licensing costs while providing the same enterprise features
- Multi-AZ deployments reduce the need for separate disaster recovery infrastructure
- The ability to scale storage and performance independently allows customers to optimize costs by provisioning only the performance needed for their workloads while leveraging unlimited elastic capacity for data growth

Service Name

Amazon FSx for OpenZFS

Service Overview

Amazon FSx for OpenZFS is a fully managed file storage service that enables users to move data from on-premises ZFS or other Linux-based file servers to AWS without changing application code or data management methods. Built on the open-source OpenZFS file system, it offers highly reliable, scalable, and feature-rich file storage with support for Linux, Windows, and macOS compute instances and containers via the NFS protocol. The service delivers up to 2 million IOPS and 21 GBps of throughput while providing powerful OpenZFS data management capabilities including data compression, snapshots, cloning, multiple volumes, thin provisioning, and user/group quotas.

Key Use Cases

- Media and Entertainment: Supporting high-performance workloads for video editing, rendering, and content distribution with low-latency access to large datasets
- Scientific Research and Analytics: Enabling genomics, financial data analytics, seismic imagery analysis, and machine learning workloads with high-throughput data processing capabilities
- Software Development and DevOps: Providing containerized applications on Amazon EKS with shared storage for development environments, testing, and CI/CD pipelines with instant snapshots and clones
- Database Applications: Supporting high-performance database workloads like Oracle with customizable record sizes and compression for optimized storage performance
- Enterprise File Sharing: Migrating on-premises ZFS or Linux-based file servers to AWS while maintaining familiar OpenZFS features and management capabilities

Features

- **OpenZFS Data Management**: Transparent compression, continuous integrity verification, snapshots, thin provisioning, and copy-on-write capabilities
- **High Performance**: Up to 2 million IOPS and 21 GBps throughput with sub-millisecond latencies for demanding workloads
- **Multiple Deployment Options**: Multi-AZ (HA), Single-AZ (HA), and Single-AZ (non-HA) deployments with different availability and durability levels
- **Intelligent-Tiering Storage**: Automatically moves data between Frequent Access, Infrequent Access, and Archive tiers based on access patterns with up to 85% cost savings

- **Multi-Protocol Access**: Supports NFS protocol access from Linux, Windows, and macOS clients, with S3 Access Points for object-style access
- **Volume Management**: Multiple volumes per file system with independent storage capacity, compression, NFS exports, and user/group quotas
- **Data Protection**: Automatic daily backups, point-in-time snapshots, on-demand replication across regions/accounts, and AWS Backup integration
- **Security and Compliance**: Encryption at rest and in transit, VPC isolation, IAM integration, and compliance with ISO, PCI DSS, SOC, and HIPAA

Value Proposition

Amazon FSx for OpenZFS provides a compelling value proposition for organizations migrating from on-premises ZFS environments or seeking high-performance NAS storage in the cloud. Unlike traditional file storage solutions, it eliminates the operational overhead of managing file servers, storage volumes, and software patching while preserving familiar OpenZFS features:

- The service offers superior cost optimization through Intelligent-Tiering (up to 85% savings compared to SSD storage), elastic scaling without upfront provisioning, and pay-as-you-go pricing
- With enterprise-grade features like instant snapshots, zero-copy clones, on-demand replication, and sub-millisecond latencies, it delivers both performance and reliability that surpasses many on-premises deployments while providing seamless AWS integration

Service Integration

Amazon FSx for OpenZFS integrates deeply with core AWS services to provide comprehensive cloud solutions. It works natively with Amazon EC2 for compute workloads, Amazon EKS and ECS for containerized applications via CSI drivers, and Amazon S3 through S3 Access Points for object-style data access:

- The service integrates with AWS Backup for centralized backup management across AWS resources, AWS CloudTrail for API logging and compliance, and Amazon CloudWatch for performance monitoring and alerting
- Additional integrations include AWS DataSync for data migration, AWS RAM for cross-account resource sharing, AWS KMS for encryption key management, Amazon VPC for network isolation, and IAM for access control
- These integrations enable customers to build comprehensive, secure, and scalable architectures

Onboarding and Offboarding

Getting started with Amazon FSx for OpenZFS is straightforward through multiple pathways. Customers can use the AWS Management Console's Quick create option for rapid deployment with default configurations, or Standard create for customized setups:

- The process involves setting up an AWS account, creating a file system in their VPC with appropriate security groups and subnets, and mounting the file system from EC2 instances using standard NFS protocols
- AWS provides comprehensive documentation, tutorials, and step-by-step guides
- For migration, customers can use AWS DataSync to transfer existing data from on-premises ZFS systems
- The service supports both programmatic access via AWS CLI/APIs and infrastructure-as-code deployment through CloudFormation templates for automated provisioning

Getting Started Guide: https://docs.aws.amazon.com/fsx/latest/OpenZFSGuide/getting-started.html

Data Removal

When offboarding from Amazon FSx for OpenZFS, customers must first terminate EC2 instances accessing the file system, then delete the file system through the AWS Console, CLI, or API. All associated backups (both automatic and user-initiated) must be explicitly deleted, as they persist independently of the file system. For data migration before deletion, customers can use AWS DataSync, standard file copy operations, or restore from backups to alternative storage. Once deleted, file system data cannot be recovered unless backups exist, emphasizing the importance of data migration planning during offboarding.

Backup/Restore and Disaster Recovery

Amazon FSx for OpenZFS provides comprehensive backup and disaster recovery capabilities through multiple mechanisms. Built-in automatic daily backups are retained up to 90 days with incremental, crash-consistent snapshots. User-initiated backups can be retained indefinitely and copied across regions for disaster recovery. Point-in-time volume snapshots enable file-level recovery and instant clone creation. On-demand data replication supports cross-region and cross-account data synchronization for disaster recovery scenarios. The service integrates with AWS Backup for centralized backup management and supports AWS Backup's logically air-gapped vaults for enhanced data protection.

Backup Capabilities

Amazon FSx for OpenZFS includes robust native backup capabilities with automatic daily backups during configurable backup windows, retained for up to 90 days. Users can create

on-demand backups retained indefinitely, with both backup types being incremental and crash-consistent:

- The service fully integrates with AWS Backup, allowing centralized backup policy management across multiple AWS services
- Backups can be copied across regions for disaster recovery, and customers can restore entire file systems from any available backup
- Additionally, point-in-time snapshots provide granular recovery options at the volume level

Disaster Recovery

Amazon FSx for OpenZFS supports comprehensive disaster recovery through multiple deployment options and data replication capabilities. Multi-AZ deployments provide automatic failover within 60 seconds across Availability Zones:

- On-demand data replication enables cross-region and cross-account disaster recovery with snapshot-based data transfer
- The service supports backup copying across regions and integrates with AWS Backup for disaster recovery orchestration
- While not directly integrated with AWS Elastic Disaster Recovery, customers can implement disaster recovery strategies using backup restoration, on-demand replication, and Multi-AZ deployments for business continuity

Recovery Objectives

Amazon FSx for OpenZFS helps customers achieve aggressive RPO and RTO objectives through multiple mechanisms. Multi-AZ deployments provide RTO under 60 seconds with synchronous replication for near-zero RPO. Point-in-time snapshots enable RPO measured in minutes to hours depending on snapshot frequency. Cross-region backup copying and on-demand replication support disaster recovery scenarios with configurable RPO/RTO based on replication frequency. Automatic daily backups provide baseline recovery capabilities, while instant snapshots and clones enable rapid recovery for testing and development scenarios.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/fsx/latest/OpenZFSGuide/limits.html

FAQ: https://aws.amazon.com/fsx/openzfs/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/fsx/latest/OpenZFSGuide/what-is-fsx.html

User Guide: https://docs.aws.amazon.com/fsx/latest/OpenZFSGuide/

API Reference: https://docs.aws.amazon.com/fsx/latest/APIReference/

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/fsx/openzfs/pricing/

Cost Optimization

Amazon FSx for OpenZFS offers unique cost optimization features centered around Intelligent-Tiering storage, which automatically moves data between three access tiers (Frequent, Infrequent, Archive) based on usage patterns, delivering up to 85% cost savings compared to SSD storage. The service eliminates upfront hardware investments and provides pay-as-you-go pricing with no minimum commitments:

- Built-in compression reduces storage consumption, while elastic storage scaling with Intelligent-Tiering means customers only pay for data stored, not provisioned capacity
- Optional SSD read cache provides performance optimization without full SSD storage costs
- Flexible throughput capacity allows right-sizing performance to workload requirements, and efficient snapshot and cloning capabilities minimize storage overhead for development and testing environments

Savings Considerations

Primary cost savings come from eliminating on-premises hardware refresh cycles, reducing operational overhead through fully managed service delivery, and leveraging Intelligent-Tiering's automatic cost optimization. Customers save up to 20% compared to traditional on-premises HDD deployments and dramatically more versus SSD storage through intelligent data tiering:

- The service eliminates costs associated with maintaining file server infrastructure, software licensing, and dedicated storage administration
- Efficient data compression, thin provisioning, and instant snapshots reduce storage consumption
- Cross-region replication and backup capabilities provide disaster recovery at lower costs than traditional solutions
- The elastic scaling model ensures customers avoid over-provisioning costs while maintaining performance during peak demand periods

Service Name

Amazon FSx for Windows File Server

Service Overview

Amazon FSx for Windows File Server is a fully managed Microsoft Windows file server service that provides enterprise-level features, performance, and compatibility for Windows workloads. Built on Windows Server, it delivers shared file storage accessible over the industry-standard Server Message Block (SMB) protocol. The service offers high levels of throughput and IOPS with consistent sub-millisecond latencies, supporting thousands of concurrent compute instances and devices. It integrates seamlessly with Microsoft Active Directory for authentication and access control, while providing administrative features like user quotas, end-user file restore, data deduplication, and File Server Resource Manager capabilities.

Key Use Cases

- Business applications: Enterprise Windows applications requiring shared file storage with Active Directory integration
- Home directories and user shares: Centralized storage for user profiles and personal folders with quota management
- High availability SQL Server: Continuously available file shares for SQL Server Always On Failover Cluster Instances
- Content management and web serving: Shared storage for content management systems and web applications
- Virtual desktop infrastructure (VDI): File storage for Amazon WorkSpaces and AppStream 2.0 environments
- Software development: Build environments and source code repositories with version control integration

Features

- **Microsoft Active Directory Integration**: Native integration with AWS Managed Microsoft AD and self-managed Active Directory for authentication and access control
- **Multi-AZ Deployment**: High availability across multiple Availability Zones with automatic failover and standby file servers
- **Data Deduplication**: Automatic identification and elimination of duplicate data, delivering typical savings of 50-60% for general-purpose file shares
- **File Server Resource Manager**: Advanced file management including quotas, file screening, classification, and storage reporting capabilities

- **Shadow Copies**: Point-in-time snapshots enabling end-user file restore and version recovery
- **Performance Scaling**: Independent scaling of storage capacity, throughput capacity, and SSD IOPS with up to 12 GB/s throughput and 400,000 IOPS
- **Hybrid Connectivity**: On-premises access via AWS Direct Connect, Site-to-Site VPN, and FSx File Gateway with local caching

Value Proposition

Amazon FSx for Windows File Server eliminates the complexity and overhead of managing Windows file server infrastructure while providing enterprise-grade features and performance. Customers benefit from fully managed operations including patching, backups, and monitoring, reducing operational burden by up to 80%:

- The service delivers native Windows compatibility with SMB protocol support, ensuring seamless integration with existing Windows applications without code changes
- Cost optimization features like data deduplication, flexible storage types (SSD/HDD), and pay-as-you-go pricing provide significant savings compared to on-premises solutions
- Built-in security features including encryption at rest and in transit, integration with AWS CloudTrail for auditing, and granular access controls meet enterprise security requirements while maintaining high availability through Multi-AZ deployments with 99.99% SLA

Service Integration

Amazon FSx for Windows File Server integrates deeply with core AWS services to provide comprehensive enterprise file storage solutions. It seamlessly works with Amazon EC2 for compute workloads, Amazon ECS for containerized applications, and VMware Cloud on AWS for hybrid environments:

- Integration with AWS Directory Service provides managed Active Directory capabilities, while AWS Backup offers centralized backup management across multiple AWS services
- The service connects with Amazon VPC for network isolation and security groups for access control
- AWS DataSync enables efficient data migration and replication, while AWS Direct Connect and Site-to-Site VPN facilitate hybrid connectivity
- Integration with Amazon CloudWatch provides monitoring and metrics, AWS CloudTrail enables audit logging, and AWS Identity and Access Management (IAM) controls administrative access

Onboarding and Offboarding

Customers can get started with Amazon FSx for Windows File Server through the AWS Management Console, AWS CLI, or API. The process involves creating an AWS account, setting up Microsoft Active Directory (either AWS Managed or self-managed), launching Amazon EC2 instances, and creating the FSx file system through the console wizard:

- The getting started guide provides step-by-step instructions including joining instances to the directory, configuring security groups, mapping file shares, and testing connectivity
- Prerequisites include an AWS account, VPC configuration, and Active Directory setup
- The service offers tutorials and documentation for common scenarios including SQL Server integration, WorkSpaces deployment, and hybrid connectivity setup
- Initial file system creation typically takes 30-45 minutes with immediate access to administrative features

Getting Started Guide: https://docs.aws.amazon.com/fsx/latest/WindowsGuide/getting-started.html

Data Removal

Data removal requires deleting the Amazon FSx file system through the AWS console, CLI, or API. Before deletion, customers should back up important data and update applications to remove file system dependencies. The deletion process permanently removes all data and cannot be undone. Customers can optionally create final backups before deletion, which can be retained independently. Cross-region backup copies provide additional protection. The process includes removing DNS aliases, updating security group references, and cleaning up any associated resources like Directory Service connections.

Backup/Restore and Disaster Recovery

Amazon FSx provides comprehensive backup and disaster recovery through automatic daily backups stored in Amazon S3 with 99.999999999% durability. User-initiated backups can be created anytime and retained indefinitely. AWS Backup integration enables centralized backup management across multiple AWS services. Cross-region and cross-account backup copying supports disaster recovery strategies. Backups are file-system-consistent, incremental, and encrypted. Restoration creates new file systems populated with backup data through lazy-loading for faster recovery times.

Backup Capabilities

Amazon FSx offers native backup capabilities with automatic daily backups enabled by default during a configurable backup window. Backups are incremental, file-system-consistent, and stored durably in Amazon S3:

- Users can create additional manual backups anytime through the console, CLI, or API
- AWS Backup integration provides centralized backup management with advanced features like cross-region copying, lifecycle management, and compliance reporting
- Backup retention can be configured from 0-90 days for automatic backups, while user-initiated backups are retained indefinitely until manually deleted

Disaster Recovery

Amazon FSx supports disaster recovery through Multi-AZ deployments providing automatic failover within regions, typically completing in under 30 seconds. Cross-region disaster recovery is achieved through backup copying and AWS DataSync replication:

- The service integrates with AWS Backup for centralized disaster recovery planning across multiple services
- Multi-AZ file systems maintain synchronous replication between primary and standby file servers
- Cross-region replication using native Microsoft DFS Replication or AWS DataSync enables geographic disaster recovery with customizable recovery point objectives

Recovery Objectives

Amazon FSx helps customers meet RPO and RTO objectives through multiple recovery mechanisms. Multi-AZ deployments provide near-zero RPO with synchronous replication and RTO under 30 seconds for automatic failover. Daily backups enable point-in-time recovery with RPO up to 24 hours. More frequent user-initiated backups can reduce RPO to minutes. Cross-region backup copying extends disaster recovery capabilities. File system restoration from backups typically matches new file system creation time, with lazy-loading reducing initial RTO for data access.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/fsx/latest/WindowsGuide/limits.html

FAQ: https://aws.amazon.com/fsx/windows/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/fsx/latest/WindowsGuide/what-is.html

User Guide: https://docs.aws.amazon.com/fsx/latest/WindowsGuide/getting-started.html

API Reference: https://docs.aws.amazon.com/fsx/latest/APIReference/welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/fsx/windows/pricing/

Cost Optimization

Amazon FSx offers several unique cost optimization features including data deduplication delivering typical savings of 30-80% through automatic duplicate data elimination. Customers can choose between SSD and HDD storage types based on performance requirements, with HDD providing significant cost savings for less performance-sensitive workloads:

- Flexible throughput capacity scaling allows independent adjustment of performance and cost
- Storage capacity can be increased as needed, avoiding over-provisioning
- Single-AZ deployments provide approximately 40% cost savings compared to Multi-AZ for non-critical workloads
- User and folder-level quotas help control storage consumption and costs
- The pay-as-you-go model eliminates upfront hardware investments and licensing fees

Savings Considerations

Cost savings come from eliminating on-premises infrastructure, reducing operational overhead by up to 80%, and leveraging AWS economies of scale. Data deduplication typically reduces storage costs by 50-60% for general-purpose file shares:

- HDD storage costs significantly less than SSD while meeting performance requirements for many workloads
- Single-AZ deployments offer substantial savings for development and non-critical environments
- Cross-region data transfer costs should be considered for disaster recovery scenarios
- AWS Backup provides cost-effective long-term retention compared to traditional backup solutions
- The managed service model reduces staffing costs for Windows file server administration, patching, and maintenance
- Integration with other AWS services can provide additional cost efficiencies through consolidated billing and volume discounts

Service Name

Amazon GameLift

Service Overview

Amazon GameLift is a fully managed service for deploying, operating, and scaling dedicated game servers for session-based multiplayer games. Built on AWS global computing infrastructure, GameLift helps deliver high-performance, high-reliability, low-cost game servers while dynamically scaling resource usage to meet worldwide player demand. The service provides automated scaling, patching, and fleet management capabilities, allowing game developers to focus on game development instead of server infrastructure management.

Key Use Cases

- Session-based multiplayer games: Hosting dedicated game servers for competitive shooters, battle royale, card games, and fighting games with low latency
- Global game deployment: Deploying game servers worldwide across AWS Regions and Local Zones to serve players with optimal connection experiences
- Auto-scaling game hosting: Managing fluctuating player demand with automatic capacity scaling and cost optimization using Spot instances

Features

- **Managed Hosting**: Fully managed EC2 and container-based game server hosting with automated deployment, scaling, patching, and fleet management
- **GameLift Anywhere**: Hybrid hosting solution allowing integration of on-premises hardware or other cloud providers with GameLift's session management
- **FlexMatch Matchmaking**: Customizable matchmaking service that automatically connects players based on latency, skill level, and game session attributes
- **Auto Scaling**: Target-based auto scaling that adjusts fleet capacity based on player demand while maintaining cost efficiency
- **Multi-Location Deployment**: Deploy game servers across multiple AWS Regions and Local Zones for global coverage and optimal player experiences
- **Game Session Management**: Comprehensive session lifecycle management including player validation, reconnection support, and session protection

Value Proposition

Amazon GameLift offers significant advantages over self-managed or alternative hosting solutions. It provides up to 70% cost savings compared to on-premises deployments through Spot instances and Graviton processors, while eliminating the complexity of infrastructure management:

- The service delivers enterprise-grade security with DDoS protection, global scale with deployment across AWS Regions and Local Zones, and built-in tools for monitoring, debugging, and matchmaking
- GameLift's deep AWS integration enables seamless connectivity with other services like CloudWatch, SNS, and DynamoDB, while supporting popular game engines including Unity and Unreal Engine through dedicated plugins

Service Integration

Amazon GameLift integrates deeply with core AWS services to provide comprehensive game hosting solutions. It leverages Amazon EC2 for compute infrastructure and Amazon ECS for container deployment, while using Amazon CloudWatch for monitoring, logging, and alarms:

- The service integrates with Amazon SNS and Amazon EventBridge for event notifications, Amazon DynamoDB for player data storage, and Amazon S3 for build artifacts and logs
- Security integrations include AWS Shield for DDoS protection, AWS IAM for access control, and Amazon Cognito for player authentication
- Additional integrations support AWS Lambda for backend processing, Amazon API Gateway for client APIs, and AWS Systems Manager for remote server access

Onboarding and Offboarding

Customers can get started with Amazon GameLift through multiple pathways designed for different expertise levels. The quickest approach uses the game server wrapper for proof-of-concept development, while Unity and Unreal Engine plugins provide guided integration workflows:

- The managed containers starter kit offers automated build and deployment solutions
- Customers must first create an AWS account, set up IAM permissions, and choose deployment regions
- The recommended learning path involves starting small with local development using GameLift Anywhere, understanding core concepts through documentation, building and testing with the preferred integration method, then scaling and optimizing for production deployment

Getting Started Guide:
https://docs.aws.amazon.com/gameliftservers/latest/developerguide/getting-started-intro.html

Data Removal

GameLift data removal involves terminating fleets, deleting builds and scripts, and removing game session logs. Customers can terminate active game sessions manually

through the console or programmatically via APIs. Fleet termination stops all instances and removes associated compute resources. Build artifacts stored in GameLift's catalog are deleted when builds are removed, while custom data stored in integrated AWS services like DynamoDB or S3 must be managed separately according to each service's deletion procedures.

Backup/Restore and Disaster Recovery

Amazon GameLift provides disaster recovery capabilities through multi-region deployment and automated failover mechanisms. Game session placement queues can be configured with multiple fleet destinations across different regions for automatic failover. The service maintains game session logs in CloudWatch and S3 for recovery purposes. However, GameLift does not provide direct backup services for game world state data, requiring customers to implement custom backup strategies using integrated AWS services.

Backup Capabilities

GameLift itself does not integrate with AWS Backup service. The service maintains game session logs and fleet configurations, but persistent game data backup must be implemented using external AWS services like S3 for file storage or DynamoDB for structured data. Customers are responsible for implementing backup strategies for game world state and player progress data through integrated services.

Disaster Recovery

GameLift does not directly integrate with AWS Elastic Disaster Recovery service. However, it provides disaster recovery capabilities through its multi-location fleet deployment and game session placement queues that can automatically failover to secondary regions. The service's built-in redundancy across AWS Regions and availability zones provides infrastructure-level disaster recovery without requiring additional DR services.

Recovery Objectives

GameLift helps meet RPO and RTO objectives through multi-region deployment, automated failover via placement queues, and real-time session management. Game session placement can failover to secondary locations within seconds, supporting low RTO requirements. For RPO objectives, customers must implement data persistence strategies using integrated AWS services, as GameLift focuses on session management rather than long-term data storage.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/general/latest/gr/gamelift.html

FAQ: https://aws.amazon.com/gamelift/faq/

Technical Requirements

Service Documentation:
https://docs.aws.amazon.com/gameliftservers/latest/developerguide/gamelift-intro.html

User Guide:
https://docs.aws.amazon.com/gameliftservers/latest/developerguide/index.html

API Reference:
https://docs.aws.amazon.com/gameliftservers/latest/apireference/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/gamelift/pricing/

Cost Optimization

GameLift offers multiple unique cost optimization strategies specifically designed for game hosting workloads. Spot instances provide up to 85% cost savings compared to On-Demand pricing, while AWS Graviton instances offer improved price-performance ratios:

- The service supports auto-scaling with configurable capacity buffers to prevent over-provisioning, and fleet replacement strategies for maintaining cost-effective AMI versions
- Container packing optimizes resource utilization by running multiple game server processes per instance
- Multi-region placement queues enable cost-optimized session placement based on both latency and hosting costs, while queue priorities can be configured to prefer lower-cost locations

Savings Considerations

Primary cost savings opportunities in GameLift include leveraging Spot instances for up to 85% savings on compute costs, implementing proper auto-scaling policies to avoid over-provisioning during low-demand periods, and using ARM-based Graviton instances for better price-performance. Container deployment enables higher resource utilization through better packing efficiency:

- Geographic cost optimization through multi-region strategies allows routing to lower-cost locations when latency permits
- The service's pay-per-use model eliminates upfront infrastructure investments, while managed services reduce operational overhead costs compared to self-managed hosting solutions

Service Name

Amazon GameLift Servers

Service Overview

Amazon GameLift Servers is a fully managed service for deploying, operating, and scaling dedicated game servers for session-based multiplayer games in the cloud. Built on AWS global computing infrastructure, it delivers high-performance, high-reliability game servers while dynamically scaling resource usage to meet worldwide player demand. The service handles server infrastructure provisioning, scaling, session management, and provides comprehensive monitoring capabilities. Amazon GameLift Servers offers flexible hosting options including managed EC2 hosting, managed container hosting, self-managed Anywhere hosting, and hybrid solutions to accommodate different deployment models and existing infrastructure.

Key Use Cases

- Session-based multiplayer games: First-person shooters, MOBAs, fighting, racing, and sports games that have defined game sessions with start and end times
- Global game server deployment: Deploying game servers across multiple AWS regions and Local Zones worldwide to minimize player latency and improve game experience
- Cost optimization for game hosting: Using Spot instances, auto-scaling, and hybrid topologies to reduce infrastructure costs while maintaining performance

Features

- **Global Deployment**: Deploy hosting in AWS Regions and Local Zones worldwide with multi-location placement queues for optimal player experiences
- **Automatic Scaling**: Target-based auto scaling with capacity buffers to handle player demand fluctuations and sudden influxes
- **FlexMatch Matchmaking**: Advanced matchmaking service supporting 2-200 concurrent players with custom rules, team-based matches, and backfilling
- **Game Session Management**: Automated game server process management, session placement, and player session validation with protection features
- **Comprehensive Monitoring**: Real-time metrics, CloudWatch integration, server logs, event tracking, and performance analysis tools
- **Flexible Hosting Options**: Managed EC2, managed containers, Anywhere hosting, and hybrid solutions with Linux/Windows support
- **Cost Optimization**: Spot instances with up to 85% savings, Graviton instances, and intelligent placement algorithms

- **Security & Reliability**: DDoS protection, TLS encryption, VPC integration, and enterprise-grade security features

Value Proposition

Amazon GameLift Servers enables developers to focus on game development rather than server infrastructure management by providing a fully managed hosting solution that can scale to support 100 million concurrent players. The service offers significant cost savings of up to 70% compared to on-premises deployments through features like Spot instances, auto-scaling, and intelligent resource allocation:

- It delivers enterprise-grade reliability and security with global reach across AWS regions and Local Zones, ensuring low-latency gameplay for players worldwide
- The service integrates seamlessly with popular game engines like Unity and Unreal Engine, and supports various deployment models from fully managed to self-managed hosting, providing flexibility for different organizational needs and existing infrastructure

Service Integration

Amazon GameLift Servers deeply integrates with the AWS ecosystem for comprehensive game hosting solutions. Core integrations include Amazon EC2 for compute infrastructure, Amazon ECS and EKS for containerized deployments, and Amazon CloudWatch for monitoring and alerts:

- The service leverages Amazon DynamoDB and Amazon S3 for game state persistence and data storage, Amazon Cognito for player authentication, and AWS Lambda with Amazon API Gateway for custom game features
- Additional integrations include Amazon Kinesis for game analytics, Amazon Chime SDK for voice chat capabilities, Amazon Elastic Load Balancing for traffic distribution, and AWS CloudFormation for infrastructure as code deployments
- The service also utilizes AWS Identity and Access Management (IAM) for security and access control, and AWS Shield for DDoS protection

Onboarding and Offboarding

Customers can get started with Amazon GameLift Servers through multiple pathways designed for different experience levels and requirements. The quickest approach uses the game server wrapper for rapid proof-of-concept development, while more comprehensive options include Unity and Unreal Engine plugins with guided workflows, and starter kits for managed containers:

- The service offers a free tier with 3,000 game sessions and 500,000 server connection minutes monthly for 12 months
- New users must create an AWS account, set up IAM permissions, and select their preferred region

- The recommended learning path involves starting small with sample games, understanding core concepts, choosing an architecture (managed EC2, containers, Anywhere, or hybrid), building and testing the integration, then scaling and optimizing for production deployment

Getting Started Guide:
https://docs.aws.amazon.com/gameliftservers/latest/developerguide/getting-started-intro.html

Data Removal

Amazon GameLift Servers provides straightforward data removal processes during offboarding. Game session logs and performance metrics are automatically retained according to configured retention policies. Build catalog data including game server builds and scripts can be deleted through the console or API. Fleet instances and associated data are removed when fleets are deleted, with no persistent data storage on individual instances. Customer-managed persistent game data stored in integrated AWS services like DynamoDB or S3 must be separately managed and deleted according to those services' data removal procedures.

Backup/Restore and Disaster Recovery

Amazon GameLift Servers inherits AWS infrastructure resilience with built-in redundancy across multiple Availability Zones within regions. The service automatically creates Point-in-Time snapshots of EBS volumes every 10 minutes for the last hour, hourly for 24 hours, and daily for 7 days. Game session data and player progress persistence depends on integration with external AWS services like DynamoDB and S3 which provide their own backup capabilities. Fleet configurations and build artifacts are stored redundantly across AWS infrastructure, ensuring availability during infrastructure failures.

Backup Capabilities

Amazon GameLift Servers does not integrate directly with AWS Backup service as it focuses on session-based gameplay rather than persistent data storage. However, individual fleet instances include 50GB EBS General Purpose SSD volumes with automatic snapshot capabilities:

- The build catalog stores game server builds redundantly across AWS infrastructure
- Persistent game data backup depends on integration with AWS Backup-supported services like DynamoDB, S3, and RDS for game state persistence, player profiles, and analytics data

Disaster Recovery

Amazon GameLift Servers does not directly integrate with AWS Elastic Disaster Recovery as it manages ephemeral game sessions rather than persistent infrastructure. The service

provides inherent disaster recovery through multi-region deployment capabilities and automatic failover mechanisms:

- Game hosting fleets can be deployed across multiple AWS regions and Local Zones, with game session placement queues automatically routing players to available resources during outages
- Fleet replacement and automated deployments enable rapid recovery of hosting capacity

Recovery Objectives

Amazon GameLift Servers helps achieve low RTO objectives through automatic scaling and multi-region deployment capabilities that can rapidly restore game hosting capacity. Game session queues provide automatic failover to alternative fleets and regions within seconds. RPO objectives are primarily relevant for persistent game data stored in integrated services like DynamoDB and S3, which provide their own recovery mechanisms. The service's session-based nature means most game data is ephemeral, focusing recovery on restoring hosting availability rather than data recovery.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/general/latest/gr/gamelift.html

FAQ: https://aws.amazon.com/gamelift/servers/faqs/

Technical Requirements

Service Documentation:
https://docs.aws.amazon.com/gameliftservers/latest/developerguide/index.html

User Guide:
https://docs.aws.amazon.com/gameliftservers/latest/developerguide/gamelift-intro.html

API Reference:
https://docs.aws.amazon.com/gameliftservers/latest/apireference/index.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/gamelift/servers/pricing/

Cost Optimization

Amazon GameLift Servers offers several unique cost optimization features designed specifically for game hosting workloads. Spot instances provide up to 85% cost savings

compared to On-Demand pricing, with intelligent placement algorithms that minimize interruptions during active game sessions:

- The service supports AWS Graviton instances for improved price-performance ratios and automatic scaling that can reduce capacity to zero when not in use
- Multi-region deployment strategies optimize costs by selecting the most cost-effective regions while maintaining player experience
- Container packing improves resource utilization by running multiple game server processes efficiently
- The service includes detailed cost monitoring through CloudWatch integration and AWS Cost Explorer, with specific metrics for game hosting optimization like percent idle instances and resource utilization tracking

Savings Considerations

Primary cost savings opportunities include leveraging Spot instances for up to 85% reduction in compute costs, implementing auto-scaling policies to match capacity with player demand, and using Linux instances which typically cost less than Windows alternatives. Game developers can save significantly by selecting appropriate instance types matched to their game requirements rather than over-provisioning resources:

- Multi-location fleets enable cost optimization by choosing regions with lower pricing while maintaining acceptable latency
- The service's pay-per-use model eliminates upfront infrastructure investments and long-term commitments
- Container-based deployments improve resource efficiency through better packing density
- Strategic use of minimum capacity settings prevents unnecessary scaling during low-demand periods, while maximum limits prevent runaway costs during traffic spikes or DDoS attacks

Service Name

Amazon GameLift Streams

Service Overview

Amazon GameLift Streams is a fully managed cloud-based game streaming service that enables game publishers to deliver high-definition gaming experiences directly to players' browsers without requiring game downloads or installations. The service uses AWS GPU instances and streaming technology to run games on cloud infrastructure and stream them at up to 1080p resolution and 60 fps to any device with a WebRTC-enabled browser. Publishers can deploy games in minutes with minimal modifications, leveraging AWS global infrastructure to provide low-latency gameplay experiences while maintaining full control over player relationships, branding, and business models.

Key Use Cases

- Direct distribution: Enable instant-play game distribution without app stores or lengthy downloads, reaching players directly through browsers
- Game demos and marketing: Create on-demand interactive game demos and trailers that potential customers can experience immediately
- Secure playtesting: Conduct internal and external game testing without distributing game builds, protecting intellectual property during development
- Legacy game revival: Relaunch back catalog games on modern devices without porting or rebuilding for new platforms
- Virtual worlds and metaverse: Build online interactive experiences and virtual worlds accessible through standard web browsers

Features

- **High-definition streaming**: Delivers gameplay at up to 1080p resolution and 60 frames per second with minimal latency
- **Multi-runtime support**: Supports Windows, Linux, and Proton runtime environments for broad game compatibility
- **Auto-scaling capacity**: Provides both always-on and on-demand capacity with automatic scaling based on player demand
- **Multi-location deployment**: Deploy across multiple AWS regions for optimal latency and global reach
- **Stream performance analytics**: Built-in monitoring through CloudWatch with real-time metrics and logging capabilities
- **WebRTC streaming protocol**: Uses open-standard WebRTC for browser compatibility without plugins or downloads

- **Data channel communication**: Enables bidirectional communication between streaming applications and web clients

Value Proposition

Amazon GameLift Streams eliminates infrastructure complexity while providing enterprise-grade streaming performance. Publishers maintain complete control over player relationships, branding, and monetization without managing servers or scaling concerns:

- The service offers instant market entry with existing games requiring minimal modifications, supporting major game engines including Unity, Unreal, and Godot
- Cost-effective scaling allows paying only for allocated capacity with flexible always-on and on-demand options
- AWS global infrastructure ensures low-latency delivery worldwide, while integrated monitoring and security features provide production-ready reliability from day one

Service Integration

Amazon GameLift Streams deeply integrates with core AWS infrastructure services. Amazon S3 stores game application files and logs, while EC2 G-series GPU instances power the streaming compute resources:

- Amazon CloudWatch provides comprehensive monitoring, metrics, and alarming capabilities
- AWS CloudTrail tracks API activity for security and compliance
- Amazon API Gateway and AWS Lambda enable serverless backend integrations for session management
- AWS Identity and Access Management (IAM) controls access to streaming resources
- The service leverages AWS global network backbone for optimized content delivery and can integrate with Amazon Cognito for player authentication workflows

Onboarding and Offboarding

Getting started involves five key steps: First, set up an AWS account with administrative access and download the GameLift Streams Web SDK. Second, upload game files to Amazon S3 using the default storage class:

- Third, create a GameLift Streams application by specifying the S3 URI and executable path
- Fourth, create a stream group defining hardware configuration, locations, and capacity requirements
- Finally, test the stream directly in the AWS console or integrate with the provided web client samples

- The service provides comprehensive documentation, sample applications, and best practice guides to accelerate deployment from development to production environments

Getting Started Guide:
https://docs.aws.amazon.com/gameliftstreams/latest/developerguide/streaming-process.html

Data Removal

Offboarding requires deleting stream groups first, which automatically releases all allocated compute capacity and terminates active sessions. Applications must be unlinked from all stream groups before deletion. Game files stored in customer-owned S3 buckets remain under customer control and must be deleted separately. Application logs and CloudWatch metrics follow standard AWS retention policies. The service provides clear resource dependencies to ensure complete cleanup and cost cessation upon stream group deletion.

Backup/Restore and Disaster Recovery

Amazon GameLift Streams operates as a stateless streaming service without built-in backup capabilities for game save data or persistent storage. Game applications and files are stored in Amazon S3, which provides 99.999999999% durability and cross-region replication options. Session state and progress must be managed by the game application itself, typically through external databases or cloud save services integrated by the developer.

Backup Capabilities

The service does not provide game state backup capabilities as it focuses on streaming compute resources. Application files stored in S3 inherit S3's durability and versioning features. GameLift Streams does not integrate with AWS Backup service since it manages ephemeral streaming sessions rather than persistent data requiring backup protection.

Disaster Recovery

GameLift Streams supports multi-region deployment for geographic redundancy but does not integrate with AWS Elastic Disaster Recovery. Recovery strategies involve recreating stream groups in alternate regions and updating DNS routing. The stateless nature of streaming sessions means disaster recovery focuses on infrastructure recreation rather than data recovery processes.

Recovery Objectives

Multi-region stream group deployment enables RPO of zero for application files stored in S3 and RTO measured in minutes for recreating streaming infrastructure. Always-on capacity in multiple regions can provide near-instantaneous failover for active streaming

sessions. Customers can achieve aggressive RTO targets by maintaining warm standby capacity across regions, though this increases operational costs for improved availability.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/gameliftstreams/latest/developerguide/quotas.html

FAQ: https://aws.amazon.com/gamelift/streams/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/gameliftstreams/latest/developerguide/index.html

User Guide: https://docs.aws.amazon.com/gameliftstreams/latest/developerguide/what-is-service.html

API Reference: https://docs.aws.amazon.com/gameliftstreams/latest/apireference/index.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/gamelift/streams/pricing/

Cost Optimization

GameLift Streams offers sophisticated capacity management with always-on, on-demand, and target-idle capacity configurations. Always-on capacity provides instant availability but incurs continuous costs, while on-demand capacity scales automatically with usage patterns:

- Target-idle capacity maintains a buffer for quick response while minimizing waste
- Multi-tenancy on high-performance stream classes allows multiple applications per GPU resource
- Proton runtime environment can reduce costs compared to Windows licensing
- Regional deployment optimization and proper capacity planning based on usage patterns are essential for cost control

Savings Considerations

Primary cost optimization involves rightsizing stream capacity based on actual usage patterns rather than peak demand. Using on-demand capacity for variable workloads and always-on only for baseline requirements significantly reduces costs:

- Multi-tenancy capabilities on supported stream classes can halve compute costs per stream
- Choosing appropriate stream classes based on game requirements prevents over-provisioning
- Regional selection impacts pricing, with some regions offering lower costs
- Proper application lifecycle management, including timely deletion of unused applications and stream groups, prevents unnecessary storage and compute charges

Service Name

Amazon GuardDuty

Service Overview

Amazon GuardDuty is an intelligent threat detection service that continuously monitors AWS accounts, workloads, and data for malicious activity and unauthorized behavior. It uses artificial intelligence (AI), machine learning (ML), anomaly detection, and integrated threat intelligence to identify known attack patterns and potential security issues across multiple AWS data sources including CloudTrail logs, VPC Flow Logs, DNS query logs, Amazon S3 data events, Amazon Aurora login events, AWS Backup data, and runtime activity for Amazon EKS, Amazon EC2, Amazon ECS, and serverless container workloads. GuardDuty provides fully managed, scalable threat detection with one-step deployment requiring no additional software or infrastructure to deploy and manage.

Key Use Cases

- Account-level threat detection: Detecting unauthorized access, credential compromise, and anomalous API activities across AWS accounts
- Container and Kubernetes security: Monitoring Amazon EKS clusters for container-based exploits and unauthorized access through audit logs and runtime monitoring
- Data protection and compliance: Monitoring S3 buckets for data exfiltration attempts, unauthorized policy changes, and malware in stored objects
- Database security monitoring: Analyzing Amazon Aurora and RDS login activity for potential access threats and unauthorized database activities
- Multi-stage attack detection: Using Extended Threat Detection to correlate events across multiple data sources and identify sophisticated attack sequences
- Malware detection: Scanning EC2 instances, EBS volumes, S3 objects, and AWS Backup data for malware and malicious files

Features

- **Foundational Threat Detection**: Continuously monitors CloudTrail management events, VPC Flow Logs, and DNS query logs using machine learning and threat intelligence
- **S3 Protection**: Monitors S3 buckets for potential security risks including data exfiltration attempts and unauthorized policy changes
- **EKS Protection**: Analyzes Kubernetes audit logs from Amazon EKS clusters to detect container-based exploits and unauthorized activities
- **Runtime Monitoring**: Provides deeper visibility into threats by analyzing operating system-level events on EC2 instances, EKS clusters, and container workloads

- **Malware Protection**: Scans EBS volumes, EC2 instances, S3 objects, and AWS Backup data for malware without requiring agents
- **RDS Protection**: Analyzes login activity for Amazon Aurora and RDS databases to detect potential access threats
- **Lambda Protection**: Monitors Lambda function network activity through VPC Flow Logs to detect threats in serverless environments
- **Extended Threat Detection**: Uses AI/ML to correlate security signals across data sources and identify multi-stage attack sequences with Critical severity findings

Value Proposition

GuardDuty provides comprehensive, intelligent threat detection with no upfront infrastructure investment or ongoing maintenance requirements. It offers immediate deployment across AWS environments with machine learning-powered detection optimized specifically for cloud threats:

- The service provides cost-effective pay-as-you-go pricing with a 30-day free trial, automatic scaling without performance impact, and seamless integration with other AWS security services
- GuardDuty's Extended Threat Detection capabilities use advanced AI/ML to identify sophisticated multi-stage attacks that traditional security tools might miss, while its protection plans offer specialized monitoring for containers, databases, serverless functions, and data storage with industry-leading threat intelligence from AWS and third-party sources

Service Integration

GuardDuty integrates natively with core AWS services for comprehensive security coverage. It automatically analyzes data from AWS CloudTrail, Amazon VPC Flow Logs, Route 53 DNS logs, Amazon S3, Amazon EKS, Amazon Aurora, Amazon RDS, AWS Lambda, and AWS Backup:

- For response and automation, GuardDuty integrates with Amazon EventBridge for real-time event processing, Amazon SNS for notifications, AWS Lambda for automated remediation, and Amazon CloudWatch for monitoring and alerting
- It works seamlessly with AWS Security Hub for centralized security posture management, Amazon Detective for investigation workflows, AWS Organizations for multi-account management, and AWS IAM for access control
- GuardDuty findings can be exported to Amazon S3 for long-term storage and analysis beyond the 90-day retention period

Onboarding and Offboarding

Getting started with GuardDuty requires only a few clicks in the AWS Management Console with no additional software or infrastructure deployment. Customers can enable

GuardDuty in standalone account environments by opening the GuardDuty console, selecting 'Amazon GuardDuty - All features', and clicking 'Get started' followed by 'Enable GuardDuty':

- For multi-account environments, customers can designate a GuardDuty administrator account and add member accounts through AWS Organizations or invitation-based management
- GuardDuty immediately begins monitoring foundational data sources upon activation, with Extended Threat Detection enabled by default
- Customers can generate sample findings to familiarize themselves with the interface and configure automated responses through EventBridge integration with SNS topics or Lambda functions for immediate operational readiness

Getting Started Guide:
https://docs.aws.amazon.com/guardduty/latest/ug/guardduty_settingup.html

Data Removal

Customers can suspend GuardDuty to stop monitoring and finding generation while retaining existing findings, or completely disable GuardDuty to permanently delete all findings and configurations. Before offboarding, all member accounts must be disassociated or deleted. Suspending preserves data for potential re-enablement, while disabling permanently removes all GuardDuty data and cannot be recovered. Malware Protection for S3 operates independently and requires separate management during offboarding.

Backup/Restore and Disaster Recovery

GuardDuty itself does not provide backup and disaster recovery capabilities as it is a monitoring and threat detection service. However, it enhances backup security through Malware Protection for AWS Backup, which scans backup data for malware threats. GuardDuty findings can be exported to Amazon S3 for indefinite retention beyond the default 90-day period, providing long-term security event archival.

Backup Capabilities

GuardDuty does not have traditional backup capabilities as it is a security monitoring service. However, it integrates with AWS Backup through Malware Protection for AWS Backup to scan backup data for threats. Customers can export GuardDuty findings to S3 for long-term storage and configure EventBridge integration for automated processing and retention of security events.

Disaster Recovery

GuardDuty does not support disaster recovery scenarios directly as it is a threat detection service. It operates regionally and customers must enable it in each desired region.

GuardDuty enhances overall disaster recovery security posture by monitoring backup data through AWS Backup integration and providing threat detection during recovery operations.

Recovery Objectives

GuardDuty supports recovery objectives indirectly by providing continuous security monitoring during disaster recovery operations and scanning backup data for malware threats. It helps maintain security posture during recovery processes by detecting threats in real-time, ensuring that restored systems and data are not compromised, though it does not directly contribute to RPO/RTO metrics.

Service Constraints

Service Quotas:
https://docs.aws.amazon.com/guardduty/latest/ug/guardduty_limits.html

FAQ: https://aws.amazon.com/guardduty/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/guardduty/latest/ug/

User Guide: https://docs.aws.amazon.com/guardduty/latest/ug/

API Reference: https://docs.aws.amazon.com/guardduty/latest/APIReference/

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/guardduty/pricing/

Cost Optimization

GuardDuty offers several cost optimization features including a 30-day free trial for most protection plans, pay-as-you-go pricing based on data volume analyzed, and volume-based discounts for VPC Flow Logs and DNS query log analysis. Customers can optimize costs by selectively enabling protection plans based on their workload requirements, using trusted IP lists to reduce false positives, and leveraging the usage monitoring dashboard to track spending by data source:

- Malware Protection for S3 includes a 12-month free tier, and incremental scanning reduces costs by only scanning changed data
- The service provides granular cost visibility through the GuardDuty console usage page, allowing customers to monitor estimated monthly spend broken down by individual data sources and protection plans

Savings Considerations

Primary cost savings come from the fully managed service model eliminating infrastructure and operational overhead compared to self-managed threat detection solutions. The pay-as-you-go model means customers only pay for actual data analyzed rather than fixed capacity, with significant volume discounts available:

- The 30-day free trial allows cost evaluation before commitment, and selective protection plan enablement lets customers optimize for their specific security requirements
- GuardDuty's automation capabilities reduce manual security operations costs, while its integration with other AWS services eliminates the need for separate security tools and reduces overall security stack complexity and associated licensing costs

Service Name

Amazon Inspector

Service Overview

Amazon Inspector is an automated vulnerability management service that continuously scans AWS workloads for software vulnerabilities and unintended network exposure. It automatically discovers and assesses Amazon EC2 instances, container images in Amazon ECR, AWS Lambda functions, and code repositories for known vulnerabilities and security issues. The service provides risk-based prioritization with Inspector risk scores, integrates seamlessly with AWS Security Hub and Amazon EventBridge for centralized security management, and supports both agent-based and agentless scanning methods to maximize coverage across diverse infrastructure configurations.

Key Use Cases

- Continuous vulnerability scanning: Automatically assess EC2 instances, Lambda functions, and container images for software vulnerabilities without manual intervention
- DevSecOps integration: Integrate vulnerability scanning into CI/CD pipelines to identify security issues before production deployment
- Multi-account security management: Centrally manage vulnerability assessments across multiple AWS accounts using AWS Organizations with delegated administrator capabilities
- Compliance and governance: Meet security compliance requirements by continuously monitoring workloads for vulnerabilities and generating comprehensive security reports
- Container security: Scan container images in Amazon ECR repositories for operating system packages and programming language vulnerabilities throughout the development lifecycle

Features

- **Automated Discovery and Scanning**: Continuously discovers and scans EC2 instances, Lambda functions, ECR container images, and code repositories without requiring manual configuration
- **Inspector Risk Score**: Provides contextualized risk scoring based on CVE information, network accessibility, and environment-specific factors to prioritize remediation efforts
- **Multi-Scan Methods**: Supports both agent-based scanning using SSM Agent and agentless scanning using EBS snapshots for maximum coverage flexibility

- **AWS Organizations Integration**: Enables centralized management across multiple accounts with delegated administrator capabilities and organization-wide policy enforcement
- **Security Hub Integration**: Automatically publishes findings to AWS Security Hub and Amazon EventBridge for centralized security monitoring and automated response workflows
- **Code Security Scanning**: Scans application source code, dependencies, and Infrastructure as Code for vulnerabilities using SAST, SCA, and IaC analysis
- **CIS Benchmark Assessments**: Performs Center for Internet Security benchmark assessments for operating system security configuration compliance

Value Proposition

Amazon Inspector eliminates the operational overhead of deploying and maintaining traditional vulnerability management solutions by providing fully managed, automated scanning across all AWS workloads. Unlike third-party solutions, Inspector offers native AWS integration with Security Hub, EventBridge, and Organizations, enabling seamless automation and centralized management:

- The service provides highly contextualized risk scoring that considers network accessibility and environment-specific factors, delivering more accurate prioritization than standalone CVE scanners
- With support for both agent-based and agentless scanning, Inspector maximizes coverage while minimizing infrastructure requirements
- The service scales automatically with your AWS environment and includes a 15-day free trial with pay-per-scan pricing that eliminates upfront investments

Service Integration

Amazon Inspector integrates deeply with core AWS security and management services to provide comprehensive vulnerability management. It publishes findings directly to AWS Security Hub for centralized security monitoring and to Amazon EventBridge for automated workflow triggers and notifications:

- The service leverages AWS Systems Manager Agent for agent-based scanning and uses Amazon EBS snapshots for agentless assessments
- Inspector integrates with AWS Organizations for multi-account management and policy enforcement, enabling delegated administrator capabilities
- It works seamlessly with Amazon ECR for container image scanning and AWS Lambda for serverless function assessments
- The service also integrates with developer tools for CI/CD pipeline scanning and exports Software Bill of Materials (SBOM) data for compliance reporting

Onboarding and Offboarding

Getting started with Amazon Inspector requires minimal configuration through the AWS Management Console or APIs. Users can activate Inspector with a single click, which automatically enables all scan types (EC2, ECR, Lambda, and Code Security) by default:

- The service creates necessary service-linked roles automatically and begins discovering eligible resources immediately
- For multi-account environments, organizations can designate a delegated administrator and use AWS Organizations policies for automated deployment across accounts and OUs
- Inspector starts scanning within minutes of activation and provides immediate visibility through the dashboard
- The service includes hybrid scanning by default, combining agent-based and agentless methods for maximum coverage without requiring pre-configuration of scanning infrastructure

Getting Started Guide:
https://docs.aws.amazon.com/inspector/latest/user/getting_started_tutorial.html

Data Removal

Deactivating Amazon Inspector removes all scan settings, suppression rules, and findings data from your account. Users should export findings to Amazon S3 before deactivation to retain vulnerability reports. The deactivation process can be performed through the console or API, and results in immediate cessation of scanning and billing. For organization-managed deployments, deactivation must be done through AWS Organizations policies. Inspector automatically closes and deletes findings after 3 days when resources are terminated or become ineligible for scanning.

Backup/Restore and Disaster Recovery

Amazon Inspector does not provide traditional backup and disaster recovery capabilities as it is a scanning service that generates findings rather than storing critical business data. The service operates regionally and findings are stored within the activated region. Users can export findings and SBOM data to Amazon S3 for external backup purposes. Inspector automatically recreates findings upon resource rescans, making data restoration unnecessary. The service relies on AWS infrastructure resilience and does not require separate disaster recovery planning.

Backup Capabilities

Amazon Inspector does not have built-in backup capabilities and does not integrate with AWS Backup since it is a vulnerability assessment service that generates transient findings data. Users can export findings as reports to Amazon S3 for retention purposes:

- The service automatically manages finding lifecycle, closing findings when vulnerabilities are remediated and recreating them during rescans
- Inspector stores findings regionally and does not require traditional backup mechanisms

Disaster Recovery

Amazon Inspector does not support traditional disaster recovery mechanisms and does not integrate with AWS Elastic Disaster Recovery. The service is regionally deployed and findings are specific to the region where resources are scanned:

- In case of regional failure, Inspector would need to be reactivated in alternate regions where workloads are deployed
- The scanning functionality would resume automatically once resources are available in the recovery region without requiring data restoration

Recovery Objectives

Amazon Inspector does not directly contribute to RPO and RTO objectives as it is a vulnerability assessment service rather than a business continuity solution. However, Inspector supports security resilience by continuously identifying vulnerabilities that could impact system availability. The service enables rapid security assessment of disaster recovery environments by automatically scanning recovered resources. Inspector's continuous scanning ensures that security posture is maintained during recovery scenarios, supporting overall business continuity objectives through proactive vulnerability management.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/inspector/latest/user/quotas.html

FAQ: https://aws.amazon.com/inspector/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/inspector/latest/user/index.html

User Guide: https://docs.aws.amazon.com/inspector/latest/user/what-is-inspector.html

API Reference: https://docs.aws.amazon.com/inspector/v2/APIReference/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/inspector/pricing/

Cost Optimization

Amazon Inspector offers several cost optimization strategies including resource exclusion through tagging to prevent scanning of non-critical instances, which eliminates unnecessary charges. The service provides a 15-day free trial for each scan type, allowing customers to evaluate usage before incurring costs:

- Organizations can optimize costs by using hybrid scanning mode that automatically selects the most cost-effective scanning method (agent-based vs
- agentless) based on instance characteristics
- Customers can also leverage suppression rules to filter out acceptable findings, reducing noise and focusing scanning resources on critical vulnerabilities
- The service supports selective activation of scan types, enabling customers to enable only required scanning capabilities rather than all features

Savings Considerations

The primary cost savings with Amazon Inspector come from eliminating the need for traditional vulnerability scanning infrastructure and associated operational overhead. Customers save on licensing costs for third-party vulnerability scanners, dedicated scanning servers, and specialized security staff for tool management:

- The pay-per-scan pricing model eliminates upfront capital investments and scales costs directly with usage
- Inspector's automated discovery and continuous scanning reduce the need for manual security assessments and penetration testing
- Integration with existing AWS services like Security Hub and EventBridge eliminates costs for additional integration tools
- The service's ability to automatically close remediated findings reduces ongoing storage and management costs compared to solutions that accumulate stale vulnerability data

Service Name

Amazon Interactive Video Service

Service Overview

Amazon Interactive Video Service (Amazon IVS) is a managed live-video streaming service powered by Twitch technology that enables developers to create channels, build interactive experiences, and distribute video at scale without managing infrastructure. It offers two streaming modes: Low-Latency Streaming with latency under 5 seconds for channels, and Real-Time Streaming with latency under 300ms for stages. The service handles video processing, delivery, and playback while allowing integration into websites and applications through comprehensive SDKs for web, iOS, and Android platforms.

Key Use Cases

- Interactive Live Streaming: Creating engaging experiences with real-time chat, polls, and Q&A alongside low-latency video streaming
- Real-Time Collaboration: Supporting collaborative live streams, host-vs-host events, live video auctions, and turning viewers into hosts
- Community Building: Building social communities with chat rooms, monetizing video content with e-commerce integration, and hosting collaborative audio rooms
- Gaming and Entertainment: Supporting live gaming streams, educational content delivery, and enterprise communications with interactive features
- Video-on-Demand: Recording live streams for later viewing and creating content libraries with automatic S3 storage integration

Features

- **Low-Latency Streaming**: Delivers video with latency under 5 seconds using optimized infrastructure and supports millions of viewers with automatic scaling
- **Real-Time Streaming**: Provides ultra-low latency under 300ms for up to 10,000 viewers and 12 hosts with collaborative live streams
- **Multi-Platform SDKs**: Broadcast and Player SDKs for iOS, Android, and web with support for RTMP, RTMPS, and SRT protocols
- **Interactive Chat**: Built-in chat APIs with message delivery, moderation capabilities, and integration with EventBridge for real-time interactions
- **Adaptive Bitrate Streaming**: Pre-configured transcoding with multiple quality levels and automatic bitrate adjustment based on viewer conditions
- **Recording and VOD**: Automatic recording to Amazon S3 with individual participant recording and composite recording options
- **Server-Side Composition**: Offloads composition and broadcasting of stages to IVS-managed service with customizable layouts

- **Multitrack Video**: Allows encoding multiple video quality levels from devices to reduce server-side transcoding costs by up to 75%

Value Proposition

Amazon IVS offers significant advantages over alternatives through its fully managed infrastructure that eliminates the need for custom video workflows, powered by proven Twitch technology for reliability at scale. The service provides industry-leading low latency with real-time streaming under 300ms and low-latency streaming under 5 seconds, coupled with global content delivery and automatic scaling:

- Cost-effectiveness is enhanced through pay-as-you-go pricing, multitrack video capability that reduces costs by up to 75%, and integrated AWS services ecosystem
- The comprehensive SDK support across web, iOS, and Android platforms, combined with built-in interactive features like chat and metadata injection, enables rapid development of engaging video experiences without infrastructure management complexity

Service Integration

Amazon IVS integrates deeply with core AWS services to provide a comprehensive streaming solution. Amazon S3 serves as the primary storage destination for recorded live streams and video-on-demand content with automatic recording capabilities:

- Amazon CloudFront acts as the content delivery network for global video distribution with optimized edge locations
- Amazon CloudWatch provides comprehensive monitoring, analytics, and alerting for stream health, viewer metrics, and performance indicators
- AWS Identity and Access Management (IAM) controls access and permissions for playback authorization and resource management
- Amazon EventBridge enables event-driven architectures by sending stream state changes, health updates, and composition events to trigger automated workflows
- Additional integrations include Amazon Kinesis Data Firehose for data streaming, AWS Elemental MediaLive for advanced broadcasting features, and Amazon Elastic Transcoder for additional video processing capabilities

Onboarding and Offboarding

Getting started with Amazon IVS involves creating an AWS account and setting up IAM permissions for service access. Customers begin by creating a channel through the AWS Management Console, AWS CLI, or SDKs, which provides a stream key and ingest endpoint for broadcasting:

- For low-latency streaming, users configure streaming software like OBS Studio with the provided credentials, while real-time streaming requires creating a stage and generating participant tokens for hosts and viewers
- The service offers comprehensive getting started guides for integrating player and broadcast SDKs across web, iOS, and Android platforms
- Optional configuration includes enabling recording to S3, setting up chat functionality, and configuring service quotas for expected usage levels
- The entire setup process can be completed in minutes with live streaming operational immediately after channel creation

Getting Started Guide:
https://docs.aws.amazon.com/ivs/latest/LowLatencyUserGuide/getting-started.html

Data Removal

Amazon IVS data removal involves deleting channels, stream keys, recording configurations, and associated resources through the console or APIs. Recorded content stored in Amazon S3 follows standard S3 data lifecycle policies and can be deleted using S3 operations. Real-time streaming resources including stages, participant tokens, and compositions must be individually deleted. Chat room data and message history are removed when rooms are deleted. Customers maintain full control over their data and can implement automated cleanup processes using AWS APIs and lifecycle policies.

Backup/Restore and Disaster Recovery

Amazon IVS provides limited native backup capabilities as a managed streaming service focused on live content delivery. Recorded live streams are automatically stored in Amazon S3 where customers can implement backup strategies using S3 Cross-Region Replication, versioning, and lifecycle policies. Configuration data for channels and stages can be backed up through infrastructure-as-code templates and API exports. Disaster recovery relies on the service's global infrastructure with automatic failover capabilities and regional redundancy for content delivery.

Backup Capabilities

Amazon IVS does not directly integrate with AWS Backup as it primarily handles live streaming rather than persistent data storage. However, recorded content stored in Amazon S3 can be protected using S3's native backup features and third-party backup solutions. Channel configurations and settings can be exported and version-controlled through infrastructure-as-code practices for configuration backup and recovery scenarios.

Disaster Recovery

Amazon IVS does not integrate with AWS Elastic Disaster Recovery as it is a managed service with built-in redundancy. The service operates on AWS's global infrastructure with automatic failover capabilities and regional availability. Disaster recovery planning focuses

on configuration backup through infrastructure-as-code and implementing multi-region streaming strategies where business requirements dictate geographic redundancy for content delivery.

Recovery Objectives

Amazon IVS supports recovery objectives through its managed service architecture with built-in high availability and automatic failover. For live streaming scenarios, the service provides rapid recovery through its global infrastructure and automatic scaling capabilities. RPO objectives are met through S3 recording with configurable frequency, while RTO objectives benefit from the service's immediate availability and global edge locations for content delivery resumption.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/general/latest/gr/ivs.html

FAQ: https://aws.amazon.com/ivs/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/ivs/latest/LowLatencyUserGuide/

User Guide: https://docs.aws.amazon.com/ivs/latest/LowLatencyUserGuide/

API Reference:
https://docs.aws.amazon.com/ivs/latest/LowLatencyAPIReference/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/ivs/pricing/

Cost Optimization

Amazon IVS offers several unique cost optimization opportunities including Multitrack Video capability that reduces input costs by up to 75% by enabling client-side encoding of multiple quality levels instead of server-side transcoding. The service provides tiered pricing for video output based on resolution and geographic regions, with significant cost savings for standard channels using multitrack compared to traditional transcoded channels ($0.50 vs $2.00 per hour):

- Pay-as-you-go pricing eliminates upfront commitments, and customers can optimize costs by choosing appropriate channel types (Basic, Standard, Advanced) based on their transcoding needs

- The AWS Free Tier provides 5 hours of live video input, 100 hours of SD output, and other allowances monthly for the first 12 months

Savings Considerations

Primary cost savings strategies include implementing Multitrack Video for standard channels to achieve up to 75% reduction in input costs, selecting appropriate channel types based on transcoding requirements, and leveraging the global content delivery network to minimize data transfer costs. Customers can optimize recording costs by implementing S3 lifecycle policies for stored content and using intelligent tiering for long-term storage:

- Real-time streaming costs can be managed by monitoring participant hours and implementing efficient stage management practices
- Geographic region selection and usage pattern analysis help optimize output costs through tiered pricing structures, while integrated AWS services provide economies of scale compared to standalone solutions

Service Name

Amazon Kendra

Service Overview

Amazon Kendra is a managed information retrieval and intelligent search service that uses natural language processing and advanced deep learning models for semantic and contextual search. It allows users to connect multiple data repositories to an index and ingest and crawl documents to create a unified search experience. The service uses machine learning algorithms to understand the context of queries and content, providing accurate and relevant results including specific answers to questions, not just search results. Kendra supports factoid, descriptive, and keyword-based natural language queries across enterprise content scattered across multiple locations and repositories.

Key Use Cases

- Enterprise search: Searching across internal documents, websites and applications when content is scattered across multiple locations and repositories
- Knowledge management: Building AI-powered assistants for financial planning, analysis, and data discovery to support analysts with business insights
- Customer support: Creating intelligent Q&A systems and chatbots that can provide specific answers from knowledge bases and documentation
- Retrieval Augmented Generation (RAG): Integrating with Amazon Bedrock Knowledge Bases and Amazon Q Business to create RAG-powered digital assistants

Features

- **GenAI Enterprise Edition**: Delivers highest accuracy for RAG use cases with advanced semantic models and information retrieval technologies
- **Intelligent Search**: Machine learning algorithms understand context of queries and content to provide accurate and relevant results
- **Data Source Connectors**: Native connectors for 30+ data sources including S3, SharePoint, Confluence, Salesforce, and databases
- **Experience Builder**: No-code interface to create and customize search experiences with filtering, FAQs, and sorting capabilities
- **Custom Document Enrichment**: Extract metadata and entities from documents using Amazon Comprehend for improved search accuracy
- **User Access Control**: Metadata-based user permissions filtering and integration with AWS IAM Identity Center
- **Query Suggestions**: Autocompletion suggestions and spell checking to help users refine queries

- **Search Analytics Dashboard**: Monitor search usage and performance metrics to optimize search experiences

Value Proposition

Amazon Kendra provides enterprise-grade intelligent search capabilities that eliminate the complexity of building and maintaining search infrastructure. Unlike traditional keyword-based search engines, Kendra uses advanced machine learning to understand natural language queries and provide direct answers, not just document links:

- The service offers pre-built connectors to 30+ enterprise data sources, reducing integration effort
- With features like GenAI Enterprise Edition optimized for RAG applications, seamless AWS service integration, and managed infrastructure that scales automatically, organizations can deploy production-ready intelligent search solutions without the overhead of managing search algorithms, relevance tuning, or infrastructure scaling

Service Integration

Amazon Kendra integrates deeply with core AWS services including Amazon S3 for document storage and indexing, Amazon Bedrock for generative AI capabilities and Knowledge Bases, Amazon Q Business for enterprise AI assistants, AWS IAM Identity Center for user access control, Amazon Comprehend for document enrichment and entity extraction, AWS Secrets Manager for secure credential storage, Amazon VPC for network isolation, AWS Lambda for custom processing, and Amazon CloudWatch for monitoring and logging. The service also works with AWS data services like Amazon RDS, DynamoDB, and Aurora for database content indexing, providing a comprehensive ecosystem for enterprise search and AI applications.

Onboarding and Offboarding

Customers get started by creating an AWS account and setting up an Amazon Kendra index through the AWS console, CLI, or SDKs. The process involves selecting an index type (GenAI Enterprise, Basic Enterprise, or Developer Edition), configuring data sources using pre-built connectors, and setting up IAM roles for access permissions:

- Amazon Kendra provides a free trial for GenAI Enterprise and Developer Edition indexes for the first 30 days
- Getting started guides are available for various data sources including Amazon S3, MySQL databases, and AWS IAM Identity Center, with step-by-step instructions for console, CLI, Python SDK, and Java SDK implementations

Getting Started Guide: https://docs.aws.amazon.com/kendra/latest/dg/getting-started.html

Data Removal

Amazon Kendra provides data removal capabilities through index deletion and data source management. Customers can delete individual documents, remove entire data sources, or delete complete indexes to remove all associated data. When deleting an index, all documents, metadata, and associated resources are permanently removed. Data source connectors can be configured to track deletions during sync operations, automatically removing documents from the index when they are deleted from source systems. For complete offboarding, customers delete all indexes, data sources, and associated IAM roles.

Backup/Restore and Disaster Recovery

Amazon Kendra does not provide explicit backup and disaster recovery capabilities as a managed service feature. The service relies on AWS infrastructure resilience and availability zones for high availability. Data durability is maintained through AWS's underlying storage systems. For disaster recovery, customers must implement their own backup strategies by maintaining copies of source documents in durable storage like Amazon S3, and re-indexing data in alternate regions if needed. Index configurations and metadata can be recreated using Infrastructure as Code approaches.

Backup Capabilities

Amazon Kendra does not have built-in backup capabilities and does not integrate with AWS Backup service. The managed service maintains data durability through AWS infrastructure, but customers are responsible for backing up their source content and index configurations. Organizations typically implement backup strategies at the source data level using services like AWS Backup for supported data sources, and maintain Infrastructure as Code templates to recreate index configurations and data source connections if needed.

Disaster Recovery

Amazon Kendra does not integrate with AWS Elastic Disaster Recovery service. Disaster recovery must be implemented through manual processes including maintaining source data backups, using multi-region index deployment strategies, and recreating indexes in alternate regions during outages. The service provides high availability within regions through AWS infrastructure, but cross-region disaster recovery requires customer-implemented solutions involving data replication and index recreation in secondary regions.

Recovery Objectives

Amazon Kendra supports recovery objectives primarily through high availability within AWS regions using multiple availability zones. For achieving specific RPO and RTO targets, customers must implement their own strategies including maintaining real-time data

source replication, automated index recreation processes, and traffic routing between regions. The service's managed nature provides operational resilience, but meeting strict recovery objectives requires combining Kendra with other AWS services like Route 53 for DNS failover and automated deployment pipelines.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/kendra/latest/dg/quotas.html

FAQ: https://aws.amazon.com/kendra/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/kendra/latest/dg/what-is-kendra.html

User Guide: https://docs.aws.amazon.com/kendra/latest/dg/getting-started.html

API Reference: https://docs.aws.amazon.com/kendra/latest/APIReference/welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/kendra/pricing/

Cost Optimization

Amazon Kendra offers unique cost optimization through multiple index editions with different pricing tiers. GenAI Enterprise Edition provides lower starting costs at $0.32/hour with smaller capacity units optimized for RAG applications:

- Organizations can optimize costs by selecting appropriate index types based on use cases, implementing efficient data source sync schedules to minimize connector charges, right-sizing storage and query capacity units based on actual usage patterns, and using the 30-day free tier for evaluation and testing
- Cost optimization also includes optimizing document sizes and extraction processes to stay within capacity limits

Savings Considerations

Main cost savings come from selecting the right index edition for specific needs, with GenAI Enterprise Edition offering 75% lower hourly costs compared to Basic Enterprise Edition. Organizations save through efficient connector management with flat monthly rates for GenAI Edition versus per-document scanning charges in Basic Edition:

- Additional savings include optimizing query patterns to reduce capacity unit requirements, implementing effective document lifecycle management to control storage costs, and using regional availability strategically
- The service's managed nature eliminates infrastructure and operational costs associated with self-hosted search solutions, providing overall cost efficiency for enterprise search implementations

Service Name

Amazon Keyspaces (for Apache Cassandra)

Service Overview

Amazon Keyspaces (for Apache Cassandra) is a scalable, highly available, and fully managed Apache Cassandra-compatible database service. It eliminates the need to provision, patch, or manage servers, and removes the requirement to install, maintain, or operate software. Amazon Keyspaces is serverless, allowing you to pay only for the resources you use while automatically scaling tables up and down in response to application traffic. The service supports building applications that serve thousands of requests per second with virtually unlimited throughput and storage capacity.

Key Use Cases

- Low latency applications: Process data at high speeds for applications requiring single-digit-millisecond latency, such as industrial equipment maintenance, trade monitoring, fleet management, and route optimization
- Open-source technology applications: Build applications on AWS using open-source Cassandra APIs and drivers available for multiple programming languages including Java, Python, Ruby, .NET, Node.js, PHP, C++, Perl, and Go
- Migrating Cassandra workloads to cloud: Move existing Cassandra workloads to AWS Cloud without managing infrastructure, reducing time-consuming and expensive self-management of Cassandra tables

Features

- **Point-in-Time Recovery (PITR)**: Continuous backups with restoration to any second within the last 35 days
- **Multi-Region Replication**: Replicate data across multiple AWS Regions for improved availability and resiliency
- **Change Data Capture (CDC) Streams**: Record row-level change events from tables in near-real time
- **On-Demand and Provisioned Capacity**: Flexible billing with on-demand pay-per-request or provisioned capacity modes
- **Time to Live (TTL)**: Automatically expire data from tables based on configured values
- **User-Defined Types (UDTs)**: Define data structures representing real-world data hierarchies
- **Client-Side Timestamps**: Cassandra-compatible timestamps for conflict resolution and write ordering

- **Encryption at Rest and in Transit**: Built-in security with AWS-managed or customer-managed encryption keys
- **Serverless Architecture**: Automatic scaling without infrastructure management

Value Proposition

Amazon Keyspaces eliminates the operational overhead of managing Apache Cassandra infrastructure while maintaining full compatibility with existing applications and tools. It provides a serverless experience with automatic scaling, high availability across multiple Availability Zones, and enterprise-grade security features including encryption at rest and in transit:

- The service offers significant cost advantages with recent price reductions up to 75% and flexible capacity modes that optimize costs based on usage patterns
- Organizations benefit from reduced administrative burden, improved reliability with 99.999% availability SLA, seamless integration with AWS services, and the ability to focus on application development rather than database management

Service Integration

Amazon Keyspaces integrates natively with core AWS services including Amazon S3 for data archival and analytics, Amazon EMR for big data processing, Amazon Redshift for data warehousing, AWS Lambda for serverless computing, Amazon ElastiCache for caching, AWS CloudWatch for monitoring and alerting, AWS IAM for access management and security, AWS CloudTrail for audit logging, AWS CloudFormation for infrastructure as code, and AWS PrivateLink for secure network connectivity. The service also supports integration with Kinesis Client Library (KCL) for CDC stream processing and works seamlessly with existing Apache Cassandra drivers and developer tools.

Onboarding and Offboarding

Getting started with Amazon Keyspaces requires signing up for AWS, creating an IAM user with appropriate permissions, and optionally installing cqlsh-expansion for command-line access. Customers can create keyspaces and tables using the AWS Management Console, AWS CLI, CQL editor, or AWS SDKs without deploying infrastructure or installing software:

- The service provides comprehensive tutorials, code examples in multiple programming languages (.NET, Java, Python, Kotlin), and sample applications
- AWS CloudShell is recommended for initial setup and testing
- Migration tools and best practices are available for existing Cassandra workloads

Getting Started Guide: https://docs.aws.amazon.com/keyspaces/latest/devguide/getting-started.html

Data Removal

To offboard from Amazon Keyspaces, customers can delete tables and keyspaces using the AWS Management Console, AWS CLI, or CQL commands. Point-in-time recovery allows restoration within 35 days if needed. Data is permanently removed when tables are deleted, with automatic cleanup of backups and associated resources. Multi-Region tables require deletion from all replicated regions. Customers should review and cancel any ongoing CDC streams, disable auto-scaling policies, and remove associated CloudWatch alarms before final cleanup.

Backup/Restore and Disaster Recovery

Amazon Keyspaces provides continuous backups through Point-in-Time Recovery (PITR) with restoration capabilities to any second within the last 35 days. Data is automatically replicated across multiple Availability Zones within a region for durability. Multi-Region replication enables disaster recovery across geographic regions. The service offers 99.999% availability SLA and built-in resilience without requiring additional disaster recovery services or manual backup management.

Backup Capabilities

Amazon Keyspaces includes built-in Point-in-Time Recovery (PITR) that creates continuous backups of table data automatically. PITR can be enabled per table and allows restoration to any point within the last 35 days. The service does not integrate with AWS Backup but provides native backup capabilities with automated management of backup retention and cleanup.

Disaster Recovery

Amazon Keyspaces provides disaster recovery through Multi-Region replication and automatic data replication across multiple Availability Zones. While it doesn't integrate with AWS Elastic Disaster Recovery, the service offers built-in resilience with 99.999% availability SLA and cross-region failover capabilities for comprehensive disaster recovery without additional services.

Recovery Objectives

Amazon Keyspaces helps meet RPO objectives through continuous PITR backups with point-in-time recovery to any second within 35 days, achieving near-zero RPO. RTO objectives are supported through automatic failover across Availability Zones, multi-region replication for geographic failover, and fast table restoration processes. The serverless architecture ensures rapid scaling and recovery without infrastructure provisioning delays.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/keyspaces/latest/devguide/quotas.html

FAQ: https://aws.amazon.com/keyspaces/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/keyspaces/latest/devguide/what-is-keyspaces.html

User Guide: https://docs.aws.amazon.com/keyspaces/latest/devguide/

API Reference: https://docs.aws.amazon.com/keyspaces/latest/APIReference/

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/keyspaces/pricing/

Cost Optimization

Amazon Keyspaces offers unique cost optimization through dual capacity modes: on-demand for variable workloads with automatic scaling and pay-per-request billing, and provisioned mode for predictable workloads with reserved capacity. Recent price reductions up to 75% make on-demand mode cost-effective for most provisioned workloads:

- Application Auto Scaling automatically adjusts provisioned capacity based on utilization targets
- TTL-based data expiration reduces storage costs by automatically removing expired data
- Right-sizing capacity and identifying unused resources through CloudWatch metrics optimization provides additional savings opportunities

Savings Considerations

Cost savings in Amazon Keyspaces come from eliminating infrastructure management overhead, reducing operational costs through serverless architecture, and leveraging flexible capacity modes. On-demand mode reduces costs by up to 56% for single-region and 65% for multi-region usage compared to previous pricing:

- Provisioned mode offers savings for predictable workloads with reserved capacity planning
- Additional savings include reduced TTL delete costs by 75%, elimination of hardware and software maintenance expenses, and optimized data transfer costs within AWS regions
- The service's automatic scaling prevents over-provisioning while ensuring performance requirements are met

Service Name

Amazon Kinesis Data Streams

Service Overview

Amazon Kinesis Data Streams is a fully managed, serverless streaming data service that enables real-time collection, processing, and analysis of large volumes of streaming data. It captures and processes gigabytes of data per second from various sources including website clickstreams, database event streams, financial transactions, social media feeds, IT logs, and location-tracking events. The service provides low-latency data processing with sub-second put-to-get delays, allowing applications to read data from a stream and perform real-time actions such as sending records to dashboards, generating alerts, or forwarding data to other AWS services for further processing.

Key Use Cases

- Real-time analytics: Processing streaming data for live dashboards, anomaly detection, and dynamic pricing
- Log and event data processing: Accelerated intake and processing of IT infrastructure logs, application logs, and clickstream data
- Real-time metrics and reporting: Continuous monitoring and reporting of business metrics and KPIs
- Complex stream processing: Building sophisticated data processing pipelines for machine learning and advanced analytics
- Event-driven applications: Powering applications that respond to real-time events and triggers

Features

- **Fully Managed Service**: Serverless architecture eliminates infrastructure management overhead
- **Multiple Capacity Modes**: Choose between On-demand Standard, On-demand Advantage, and provisioned modes based on throughput requirements
- **Enhanced Fan-out**: Dedicated throughput of 2 MB/sec per consumer for parallel data processing
- **Server-side Encryption**: Automatic encryption at rest using AWS KMS keys
- **High Availability**: Data synchronously replicated across three Availability Zones
- **Warm Throughput**: Pre-configured throughput capacity up to 10 GiB/s for instant scaling in On-demand Advantage mode
- **Extended Data Retention**: Configurable retention periods from 24 hours up to 365 days

- **Large Record Support**: Support for records up to 10 MiB in size

Value Proposition

Amazon Kinesis Data Streams offers compelling advantages over alternatives through its fully managed serverless architecture that eliminates operational overhead while providing enterprise-grade reliability with 99.95% availability SLA and 99.999% durability. The service delivers sub-second latency for real-time processing with automatic scaling capabilities, making it ideal for unpredictable workloads:

- Its deep integration with the AWS ecosystem enables seamless data flow to analytics services like Lambda, S3, Redshift, and EMR
- The flexible capacity modes allow organizations to optimize costs whether they have consistent or variable traffic patterns, while the enhanced fan-out capability supports multiple concurrent consumers without performance degradation

Service Integration

Kinesis Data Streams integrates extensively with AWS services as both data source and destination. Producer integrations include AWS DMS for change data capture, IoT Core for device data, CloudFront for web analytics, and Aurora/RDS for database streams:

- Consumer integrations encompass Lambda for serverless processing, Kinesis Data Firehose for data delivery to S3/Redshift/OpenSearch, Managed Service for Apache Flink for stream processing, EventBridge for event routing, SQS/SNS for messaging, and AWS Glue for ETL operations
- Analytics integrations include EMR for big data processing, Redshift for data warehousing, and QuickSight for visualization
- The service also supports third-party integrations with Apache Flink, Kafka, Spark, Databricks, and Adobe Experience Platform

Onboarding and Offboarding

Getting started with Kinesis Data Streams involves creating a stream through the AWS Console, CLI, or SDK by specifying a name and choosing between on-demand or provisioned capacity modes. For provisioned mode, determine the required number of shards based on throughput needs (1 MB/s write, 2 MB/s read per shard):

- Configure producers using the Kinesis API, Producer Library (KPL), or Kinesis Agent, and set up consumers using the Client Library (KCL), Lambda functions, or direct SDK calls
- Essential prerequisites include an AWS account, appropriate IAM permissions, and for development environments, Java 7+ with required libraries
- The service provides comprehensive tutorials and code examples for rapid implementation across multiple programming languages and frameworks

Getting Started Guide: https://docs.aws.amazon.com/streams/latest/dev/examples.html

Data Removal

Data removal is accomplished by deleting the Kinesis data stream using the DeleteStream API, AWS Console, or CLI, which permanently removes the stream, all shards, and associated data. Applications must be shut down before deletion, and the stream must be in ACTIVE state. The deletion process also removes associated tags and consumer registrations. Data is automatically purged based on the configured retention period (24 hours to 365 days), after which it becomes inaccessible. For immediate data removal, delete the stream entirely.

Backup/Restore and Disaster Recovery

Kinesis Data Streams provides built-in disaster recovery through automatic synchronous replication of data across three Availability Zones within a region. The service achieves high durability (99.999%) and availability (99.95%) through this multi-AZ architecture. Data retention periods from 24 hours to 365 days serve as a form of backup, allowing replay of historical data. For cross-region disaster recovery, customers can implement producer applications that write to streams in multiple regions or use Kinesis Data Firehose to replicate data to S3 across regions.

Backup Capabilities

Kinesis Data Streams does not integrate with AWS Backup service but provides data protection through configurable retention periods and automatic multi-AZ replication. Data is stored durably for the specified retention period (24 hours to 365 days), enabling replay and recovery scenarios:

- Extended retention allows access to historical data for up to one year
- The service's inherent design provides continuous data availability rather than traditional backup mechanisms

Disaster Recovery

The service does not integrate with AWS Elastic Disaster Recovery but provides inherent disaster recovery through multi-AZ data replication and regional resilience. Kinesis Data Streams automatically maintains copies of data across three Availability Zones, ensuring continuity during AZ failures. For broader disaster recovery strategies, customers implement cross-region replication patterns using producers that write to multiple regions or downstream services like Kinesis Data Firehose for data replication.

Recovery Objectives

Kinesis Data Streams supports aggressive RPO and RTO objectives through its real-time replication and sub-second processing capabilities. RPO approaches near-zero due to synchronous replication across Availability Zones, minimizing data loss potential. RTO is

typically minutes for AZ failures due to automatic failover mechanisms. The configurable retention period enables point-in-time recovery scenarios. For cross-region scenarios, RPO depends on replication patterns implemented by customers, while RTO involves redirecting traffic to alternate regions.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/streams/latest/dev/service-sizes-and-limits.html

FAQ: https://aws.amazon.com/kinesis/data-streams/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/streams/latest/dev/introduction.html

User Guide: https://docs.aws.amazon.com/streams/latest/dev/introduction.html

API Reference: https://docs.aws.amazon.com/kinesis/latest/APIReference/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/kinesis/data-streams/pricing/

Cost Optimization

Kinesis Data Streams offers multiple cost optimization strategies through its flexible capacity modes. On-demand Advantage mode provides up to 60% cost savings for workloads processing at least 10 MiB/s with committed usage pricing:

- Optimizing data retention periods reduces storage costs, while efficient shard management in provisioned mode prevents over-provisioning
- Data compression before ingestion reduces throughput costs
- The new warm throughput capability eliminates over-provisioning for predictable traffic spikes
- Enhanced fan-out should be used judiciously as it incurs additional per-consumer costs
- Choosing the appropriate capacity mode based on traffic patterns (on-demand for variable, provisioned for predictable) maximizes cost efficiency

Savings Considerations

Primary cost savings opportunities include selecting the optimal capacity mode for workload patterns, with on-demand modes eliminating over-provisioning costs and provisioned mode offering predictable pricing for steady workloads. On-demand Advantage mode delivers significant savings for high-throughput applications through committed usage pricing:

- Reducing retention periods from default settings can substantially lower storage costs
- Implementing data compression reduces throughput charges
- Efficient partition key strategies prevent hot spotting and reduce resharding costs
- Consolidating multiple small streams can reduce per-stream overhead costs
- Monitoring and right-sizing shard counts in provisioned mode prevents waste, while leveraging AWS Cost Explorer and billing alerts enables ongoing optimization

Service Name

Amazon Kinesis Data Streams

Service Overview

Amazon Kinesis Data Streams is a fully managed, serverless streaming data service that enables real-time collection and processing of large streams of data records. The service allows users to create data-processing applications that read streaming data from data streams as records, process them in real-time, and send results to dashboards, generate alerts, dynamically change pricing strategies, or integrate with other AWS services. It handles massive volumes of data from sources including website clickstreams, database event streams, financial transactions, social media feeds, IT logs, and location-tracking events with sub-second latency for real-time analytics and event-driven applications.

Key Use Cases

- Real-time analytics: Process streaming data for real-time dashboards, anomaly detection, and dynamic pricing strategies
- Log and event data streaming: Accelerated intake and processing of IT infrastructure logs, application logs, and system events
- Event-driven applications: Power real-time applications that respond immediately to streaming data changes and triggers
- Real-time metrics and reporting: Generate live metrics, monitoring data, and operational reports from continuous data streams
- Complex stream processing: Handle sophisticated data transformations and aggregations on high-volume streaming data
- IoT data ingestion: Process telemetry data, sensor readings, and device events from Internet of Things applications

Features

- **Serverless Operation**: Fully managed service eliminating infrastructure management overhead
- **High Availability and Durability**: Synchronously replicates data across three Availability Zones for 99.999% durability
- **Low Latency Processing**: Sub-second put-to-get delay enabling near real-time data processing
- **Multiple Capacity Modes**: On-demand Standard, On-demand Advantage, and provisioned modes for different scaling needs
- **Enhanced Fan-Out**: Dedicated 2MB/s throughput per consumer enabling parallel processing without contention

- **Server-Side Encryption**: Automatic data encryption at rest using AWS KMS with configurable keys
- **Extended Data Retention**: Configurable retention periods from 24 hours up to 365 days for long-term data access
- **Large Record Support**: Handles records up to 10 MiB for IoT, CDC pipelines, and machine learning workflows

Value Proposition

Amazon Kinesis Data Streams provides unmatched real-time streaming capabilities with enterprise-grade reliability and seamless AWS integration. The service offers automatic scaling from gigabytes to terabytes per hour without infrastructure management, while maintaining sub-second latency for mission-critical applications:

- Unlike alternatives, it provides native integration with 20+ AWS services, enhanced fan-out for multiple consumers, and flexible capacity modes including the new On-demand Advantage with 60% cost savings
- The service guarantees 99.95% availability and 99.999% durability with automatic replication across multiple Availability Zones, eliminating the operational burden of building and maintaining streaming infrastructure while ensuring enterprise-scale performance and reliability

Service Integration

Kinesis Data Streams integrates natively with essential AWS analytics and processing services for comprehensive data pipelines. Core integrations include AWS Lambda for serverless processing, Amazon Data Firehose for data delivery to destinations, Amazon EMR for big data analytics, and Amazon Managed Service for Apache Flink for stream processing:

- Additional key integrations encompass Amazon EventBridge for event routing, AWS Glue for ETL operations, Amazon Redshift for data warehousing, Amazon S3 for storage, Amazon DynamoDB for metadata management via Kinesis Client Library, and Amazon CloudWatch for monitoring
- The service also supports third-party integrations with Apache Spark, Databricks, Confluent Platform, and other popular streaming frameworks, enabling flexible architecture designs

Onboarding and Offboarding

Getting started with Amazon Kinesis Data Streams is straightforward through multiple pathways. Customers begin by creating an AWS account and accessing the Kinesis console to create their first data stream, choosing between on-demand and provisioned capacity modes based on throughput requirements:

- The service provides comprehensive tutorials including processing real-time stock data using KPL and KCL libraries, basic CLI operations, and integration with Lambda functions
- Essential tools include the AWS SDK, Kinesis Client Library (KCL), Kinesis Producer Library (KPL), and Kinesis Agent for different implementation approaches
- Customers can start with the getting started tutorials, follow the developer guide for detailed implementation, and utilize AWS documentation, code samples, and hands-on labs to build production-ready streaming applications quickly

Getting Started Guide: https://docs.aws.amazon.com/streams/latest/dev/examples.html

Data Removal

Data removal in Kinesis Data Streams occurs automatically based on the configured retention period, ranging from 24 hours to 365 days, after which data is permanently deleted and cannot be recovered. For immediate offboarding, customers can delete data streams through the console, CLI, or API, which immediately stops data ingestion and begins the deletion process. Existing data in the stream will be retained until the retention period expires, then automatically purged. Customers should ensure consumers have processed all required data before deletion, as there is no data recovery mechanism once the retention period ends.

Backup/Restore and Disaster Recovery

Amazon Kinesis Data Streams provides built-in disaster recovery through automatic data replication across three Availability Zones within a region, ensuring high durability without requiring customer-managed backups. The service achieves resilience through synchronous replication and automatic failover capabilities. For cross-region disaster recovery, customers can implement dual-streaming architectures or use Amazon Data Firehose to deliver data to durable storage like Amazon S3 for long-term backup. Data retention periods up to 365 days provide recovery windows, though traditional backup/restore mechanisms are not applicable to streaming data services.

Backup Capabilities

Kinesis Data Streams does not provide traditional backup capabilities in the conventional sense, as it is designed for real-time streaming rather than static data storage. However, it offers data durability through automatic replication across multiple Availability Zones and configurable retention periods:

- The service does not integrate with AWS Backup due to its streaming nature
- For data preservation, customers typically stream data to persistent storage services like Amazon S3 via Amazon Data Firehose or process and store data using consumer applications

Disaster Recovery

Kinesis Data Streams provides inherent disaster recovery through its architecture of synchronous data replication across three Availability Zones, enabling automatic recovery from individual zone failures. The service does not integrate with AWS Elastic Disaster Recovery as it maintains real-time data availability rather than requiring recovery procedures:

- For regional disasters, customers implement cross-region streaming strategies or use integrated services like Data Firehose to replicate data to multiple regions
- The managed service automatically handles infrastructure failures and maintains data availability

Recovery Objectives

Kinesis Data Streams supports extremely low Recovery Time Objectives (RTO) through automatic failover mechanisms that maintain continuous data availability during infrastructure failures, typically achieving RTOs of seconds to minutes. Recovery Point Objectives (RPO) align with the service's sub-second replication across Availability Zones, minimizing data loss to near-zero for most failure scenarios. The configurable retention periods up to 365 days allow customers to replay data from specific points in time, supporting various RPO requirements for downstream applications and enabling data recovery strategies.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/streams/latest/dev/service-sizes-and-limits.html

FAQ: https://aws.amazon.com/kinesis/data-streams/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/streams/latest/dev/

User Guide: https://docs.aws.amazon.com/streams/latest/dev/

API Reference: https://docs.aws.amazon.com/kinesis/latest/APIReference/

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/kinesis/data-streams/pricing/

Cost Optimization

Kinesis Data Streams offers multiple cost optimization strategies through capacity mode selection and usage patterns. On-demand Advantage mode provides 60% cost savings compared to On-demand Standard when ingesting at least 10MB/s aggregate or using multiple consumers:

- Provisioned mode offers predictable costs for consistent workloads
- Enhanced fan-out pricing is now equivalent to shared-throughput retrieval in Advantage mode, making multi-consumer architectures more cost-effective
- Extended retention costs decreased by 77% in Advantage mode
- Additional optimizations include right-sizing shard counts, optimizing partition key distribution to avoid hot sharding, reducing retention periods when possible, implementing efficient consumer applications, and consolidating multiple small streams to meet minimum throughput commitments

Savings Considerations

Primary cost savings opportunities include migrating to On-demand Advantage mode for workloads meeting the 25MB/s minimum commitment, achieving up to 60% cost reduction. Customers can optimize shard utilization in provisioned mode by monitoring throughput metrics and adjusting shard counts appropriately:

- Implementing enhanced fan-out becomes more cost-effective in Advantage mode for multiple consumer scenarios
- Data retention optimization by setting appropriate retention periods based on actual business needs rather than maximum limits provides significant savings
- Consolidating multiple low-throughput streams into fewer high-throughput streams maximizes Advantage mode benefits
- Proper partition key design prevents over-provisioning shards due to hot partition issues, and using efficient producer libraries like KPL reduces PUT payload unit charges

Service Name

Amazon Kinesis Video Streams

Service Overview

Amazon Kinesis Video Streams is a fully managed AWS service that enables secure streaming of live video and time-encoded data from connected devices to the AWS Cloud for analytics, machine learning, playback, and other processing. The service automatically provisions and scales infrastructure for ingesting streaming video data from millions of devices, then stores, encrypts, and indexes the data. It supports real-time consumption with its hot storage tier and extended retention with its warm storage tier, enabling both real-time and batch-oriented video analytics applications.

Key Use Cases

- Smart Home Applications: Streaming video from security cameras and IoT devices for real-time monitoring, motion detection, and automated alerts
- Smart City Solutions: Processing video feeds from traffic cameras, public safety systems, and urban infrastructure monitoring for analytics and decision-making
- Industrial Automation: Collecting video data from manufacturing equipment, drones, and sensors for quality control, predictive maintenance, and operational intelligence

Features

- **Secure Video Ingestion**: SDKs and APIs for streaming video data with encryption in transit using TLS and at rest using AWS KMS
- **WebRTC Support**: Low-latency two-way media streaming between web browsers, mobile applications, and connected devices
- **Video Playback**: Live and on-demand video playback via HTTP Live Streaming (HLS) and MPEG-DASH protocols
- **Amazon Rekognition Integration**: Built-in integration with Amazon Rekognition Video for face detection, object recognition, and content analysis
- **Tiered Storage**: Hot tier for immediate access and warm tier for cost-effective long-term retention with different pricing models
- **Automatic Indexing**: Time-based indexing of stored media using both producer timestamps and ingestion timestamps
- **Edge Recording**: Local recording capabilities with scheduled cloud streaming for IP cameras on customer premises

Value Proposition

Amazon Kinesis Video Streams eliminates the complexity of building and managing video streaming infrastructure by providing a fully managed, serverless solution that automatically scales from one to millions of video sources. It offers unique advantages including built-in integration with Amazon Rekognition for AI-powered video analysis, support for both real-time and batch processing workflows, and comprehensive security with encryption at rest and in transit. The service provides flexible storage tiers to optimize costs, supports diverse data types beyond video (audio, thermal, RADAR), and integrates seamlessly with the broader AWS ecosystem for comprehensive video analytics solutions.

Service Integration

Amazon Kinesis Video Streams integrates natively with Amazon Rekognition Video for face detection and object recognition, Amazon S3 as the underlying durable storage layer, and AWS Key Management Service (KMS) for encryption. It works with AWS Identity and Access Management (IAM) for access control, Amazon SNS for notifications, Amazon CloudWatch for monitoring and logging, and AWS CloudTrail for API usage tracking. The service also integrates with popular machine learning frameworks including Apache MXNet, TensorFlow, and OpenCV, enabling customers to build custom video analytics applications on Amazon EC2 instances.

Onboarding and Offboarding

Customers begin by setting up an AWS account and creating a Kinesis video stream through the AWS Management Console, CLI, or SDKs. They then configure producer devices using available SDKs for Java, C++, or Android to stream video data to the service:

- The process includes setting up proper IAM permissions, configuring stream properties like retention period and storage tier, and integrating with consuming applications
- AWS provides comprehensive documentation, sample applications, tutorials including Raspberry Pi examples, and a getting started guide to facilitate rapid onboarding and implementation

Getting Started Guide:
https://docs.aws.amazon.com/kinesisvideostreams/latest/dg/getting-started.html

Data Removal

Data removal involves deleting Kinesis video streams through the AWS Management Console, CLI, or DeleteStream API, which permanently removes all associated video data from both hot and warm storage tiers. Stream deletion is irreversible, so customers must ensure important video data is backed up before deletion. The service automatically handles cleanup of indexed metadata and associated resources when streams are deleted.

Backup/Restore and Disaster Recovery

Kinesis Video Streams provides inherent data durability through its underlying Amazon S3 storage with automatic encryption and indexing. The service supports configurable retention periods and offers warm tier storage for cost-effective long-term retention. Customers can implement additional backup strategies by integrating with other AWS services to copy video data to different regions or storage classes for disaster recovery purposes.

Backup Capabilities

The service does not integrate directly with AWS Backup but provides built-in durability through Amazon S3 storage with configurable retention periods. Customers can implement custom backup solutions by using APIs to retrieve video data and store it in additional locations or integrate with AWS services for cross-region replication.

Disaster Recovery

Kinesis Video Streams does not integrate directly with AWS Elastic Disaster Recovery but supports disaster recovery through its durable storage architecture and availability across multiple AWS regions. Customers can deploy streams in multiple regions and implement cross-region data replication strategies for comprehensive disaster recovery planning.

Recovery Objectives

The service supports low RTO through its highly available architecture across multiple Availability Zones and rapid stream recreation capabilities. For RPO objectives, configurable retention periods and durable storage ensure minimal data loss, while warm tier storage provides cost-effective long-term data retention for meeting extended recovery point requirements in disaster scenarios.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/kinesisvideostreams/latest/dg/limits.html

FAQ: https://aws.amazon.com/kinesis/video-streams/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/kinesisvideostreams/latest/dg/

User Guide: https://docs.aws.amazon.com/kinesisvideostreams/latest/dg/

API Reference:
https://docs.aws.amazon.com/kinesisvideostreams/latest/dg/API_Reference.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/kinesis/video-streams/pricing/

Cost Optimization

Kinesis Video Streams offers unique cost optimization through tiered storage with hot and warm tiers, where warm tier uses fragment-based billing instead of GB-based billing, providing significant savings for long-term retention. Customers can optimize costs by adjusting fragment duration, selecting appropriate retention periods, and choosing between storage tiers based on access patterns. The service also offers cost benefits for higher bitrate cameras in warm tier and provides flexible pricing based on actual data ingestion and consumption rather than provisioned capacity.

Savings Considerations

Primary cost savings come from the pay-as-you-use model with no upfront costs or infrastructure management overhead, elimination of the need to build and maintain streaming infrastructure, and automatic scaling that prevents over-provisioning. The warm storage tier provides up to 50% cost savings for long-term retention compared to hot tier, while integration with AWS services reduces development time and operational costs. Customers also save through reduced bandwidth costs with efficient streaming protocols and the ability to process video data without downloading entire files.

Service Name

Amazon Lex

Service Overview

Amazon Lex is a fully managed artificial intelligence service that enables developers to build, design, test, and deploy voice and text conversational interfaces in applications. Powered by the same technology as Amazon Alexa, it provides high-quality speech recognition and natural language understanding capabilities. The service utilizes Generative AI and Large Language Models (LLMs) to enhance conversational experiences by providing automated responses to frequently asked questions, analyzing customer sentiment and intents, generating summaries, and generating email or chat responses. Amazon Lex requires no deep learning expertise and offers a scalable, secure, end-to-end solution for building sophisticated chatbots for mobile devices, web applications, chat platforms, and IoT devices.

Key Use Cases

- Customer support: Build call center bots and automated customer support agents for handling inquiries and providing assistance
- E-commerce and retail: Create transactional bots for order processing, product recommendations, and inventory management
- Appointment booking: Develop scheduling bots for healthcare, services, and business appointments
- IT helpdesk: Build automated helpdesk agents for technical support and troubleshooting
- Financial services: Create bots for account inquiries, transaction processing, and financial assistance
- HR and employee services: Develop internal bots for employee queries, benefits information, and organizational support
- IVR migration: Convert traditional interactive voice response systems to conversational AI experiences

Features

- **High-Quality Speech Recognition and Natural Language Understanding**: Advanced ASR and NLU capabilities powered by the same technology as Alexa
- **Generative AI Integration**: Utilizes Large Language Models for enhanced conversational experiences, automated FAQ responses, and content generation
- **Multi-turn Dialog and Context Management**: Maintains conversation context across multiple interactions and manages complex dialog flows

- **Assisted Natural Language Understanding**: Improved intent classification and slot resolution using LLMs while staying within configured intents
- **Custom Vocabulary Support**: Supports custom vocabularies in 17 languages to improve speech recognition accuracy for specific domains
- **Visual Conversation Builder**: Drag-and-drop interface to design and visualize conversation paths and flows
- **Automated Chatbot Designer**: Design chatbots from conversation transcripts using AI-powered analysis
- **Multi-language and Multi-region Support**: Supports multiple languages in a single bot and global resiliency with bot replication across regions
- **8 kHz Telephony Audio Support**: Optimized for contact center and telephony integrations with high-quality audio processing
- **Network of Bots**: Combine multiple bots into a single network and route requests based on user input

Value Proposition

Amazon Lex provides a compelling value proposition through its integration of advanced AI technologies with enterprise-grade reliability. As the only conversational AI service powered by the same technology as Alexa, it offers proven speech recognition and natural language understanding capabilities:

- The service democratizes AI development by eliminating the need for deep learning expertise while providing access to cutting-edge features like Generative AI and Large Language Models
- Key differentiators include seamless AWS ecosystem integration, comprehensive multi-channel deployment options, built-in security and compliance features, and cost-effective pay-as-you-go pricing
- The platform's visual conversation builder, automated chatbot designer, and pre-built industry templates significantly reduce development time and complexity compared to building conversational interfaces from scratch

Service Integration

Amazon Lex integrates extensively with core AWS services to provide comprehensive conversational AI solutions. Primary integrations include AWS Lambda for backend business logic execution and intent fulfillment, Amazon Polly for neural text-to-speech capabilities, and Amazon Bedrock for advanced generative AI features through foundation models and knowledge bases:

- Contact center integrations encompass Amazon Connect, Genesys Cloud CX, Amazon Chime SDK, and AWS Contact Center Intelligence
- Additional key integrations include Amazon Kendra for intelligent search and knowledge retrieval, Amazon Comprehend for sentiment analysis, Amazon Cognito

for user authentication, Amazon DynamoDB for session management and data storage, Amazon CloudWatch for monitoring and analytics, and Amazon S3 for data storage

- These native integrations enable developers to build sophisticated, enterprise-ready conversational applications with minimal custom integration work

Onboarding and Offboarding

Getting started with Amazon Lex is straightforward through multiple pathways designed for different user types. Business users can begin with the No-Code Path using pre-built templates and the 5-minute quick start guide:

- Developers can follow the Developer Path involving template customization, Lambda integration, and API usage
- Enterprise users can pursue the Enterprise Path covering advanced features, security best practices, and global resiliency
- The process begins by signing into the AWS Management Console, navigating to Amazon Lex under Artificial Intelligence, and selecting from available bot templates or creating custom bots
- The visual conversation builder and automated chatbot designer facilitate rapid prototyping
- Prerequisites include an active AWS account, modern web browser, and appropriate IAM permissions
- Exercises typically take 30-60 minutes and cost under $1 within free tier limits

Getting Started Guide: https://docs.aws.amazon.com/lexv2/latest/dg/getting-started.html

Data Removal

Amazon Lex provides controlled data removal options for offboarding. Voice and text inputs processed by Amazon Lex are stored temporarily and can be deleted by customers through AWS console or API operations. The service architecture does not require long-term storage of end user data, and comprehensive versioning capabilities enable easy rollback. Customers retain full ownership and control over their content, with the ability to delete conversation logs, bot configurations, and associated data. The managed service architecture automatically handles resource cleanup, though customers should review and remove any integrated Lambda functions, DynamoDB tables, or other connected AWS resources as needed.

Backup/Restore and Disaster Recovery

Amazon Lex provides comprehensive backup and disaster recovery capabilities through its managed service architecture and Global Resiliency features. Bot configurations, intents, slots, and dialog flows are automatically backed up with comprehensive versioning for easy rollback. The Global Resiliency feature enables automatic bot replication, version

synchronization, and alias preservation across AWS regions in near real-time. This includes CloudFormation support for infrastructure as code and existing alias replication to maintain regional configurations. The service supports disaster recovery through multi-region deployment patterns and integrates with AWS Application Recovery Controller for automated failover management.

Backup Capabilities

Amazon Lex has built-in backup capabilities through its managed service architecture. Bot definitions, intents, slots, and conversation flows are automatically backed up with comprehensive versioning support allowing up to 100 versions per resource type:

- The service provides automated backup of bot configurations and training data without requiring customer intervention
- While Amazon Lex does not directly integrate with AWS Backup service, its managed nature ensures data protection through automated infrastructure management and version control systems that maintain historical configurations and enable rollback capabilities

Disaster Recovery

Amazon Lex supports disaster recovery through Global Resiliency features and multi-region deployment capabilities. The service does not directly integrate with AWS Elastic Disaster Recovery, but provides native disaster recovery through automatic bot replication across AWS regions, version synchronization, and alias preservation:

- Global Resiliency maintains regional configurations in sync and supports CloudFormation for infrastructure consistency
- Customers can implement disaster recovery strategies by deploying bots across multiple regions and using AWS Application Recovery Controller for failover management

Recovery Objectives

Amazon Lex helps customers meet RPO and RTO objectives through Global Resiliency features that provide near real-time bot replication and synchronization across regions. The managed service architecture minimizes data loss risk with automatic backups and versioning. For contact center applications, the service maintains regional configurations in sync, enabling rapid failover scenarios. Recovery objectives are supported through multi-region deployment patterns, automated failover capabilities with AWS Application Recovery Controller, and the ability to quickly redeploy bot configurations using CloudFormation templates.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/lexv2/latest/dg/quotas.html

FAQ: https://aws.amazon.com/lex/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/lexv2/latest/dg/what-is.html

User Guide: https://docs.aws.amazon.com/lexv2/latest/dg/getting-started.html

API Reference: https://docs.aws.amazon.com/lexv2/latest/APIReference/welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/lex/pricing/

Cost Optimization

Amazon Lex offers several unique cost optimization opportunities through its flexible pricing model and efficient resource utilization. The pay-as-you-go model charges only for actual speech ($0.004 per request) and text ($0.00075 per request) interactions, eliminating infrastructure costs:

- Streaming conversation pricing provides cost efficiency for multi-turn dialogs by processing multiple interactions in a single API call
- The automated chatbot designer charges per minute of transcript analysis ($0.50/minute), enabling cost-effective bot development from existing conversation data
- Free tier credits up to $200 for new AWS customers provide substantial cost savings during development and testing phases
- Cost optimization strategies include using text interactions where possible, implementing efficient conversation flows to minimize turn count, and leveraging pre-built templates to reduce development time and costs

Savings Considerations

Primary cost savings considerations include choosing optimal interaction models based on use case requirements, with streaming conversations offering better value for complex multi-turn dialogs compared to individual request-response patterns. Text-based interactions cost significantly less than speech processing, making hybrid approaches cost-effective:

- The managed service architecture eliminates infrastructure, maintenance, and scaling costs associated with self-hosted solutions

- Development cost savings come from pre-built industry templates, automated chatbot designer for rapid prototyping, and visual conversation builder reducing development time
- Operational savings result from automatic scaling, built-in monitoring, and comprehensive versioning eliminating custom infrastructure management
- Integration with other AWS services through native connectors reduces custom development costs, while the Global Resiliency feature provides disaster recovery without additional infrastructure investment

Service Name

Amazon Lightsail

Service Overview

Amazon Lightsail is the easiest way to get started with Amazon Web Services (AWS) for anyone who needs to build websites or web applications. It includes everything needed to launch projects quickly—instances (virtual private servers), container services, managed databases, content delivery network (CDN) distributions, load balancers, SSD-based block storage, static IP addresses, DNS management of registered domains, and resource snapshots (backups)—for a low, predictable monthly price. Lightsail simplifies cloud computing by providing pre-configured blueprints and templates, making it ideal for developers, small businesses, students, and users who prefer a simple management interface over the complexity of traditional AWS services.

Key Use Cases

- Website and web application hosting: Deploy WordPress, Drupal, or custom applications with pre-configured templates
- Development and testing environments: Create dev sandboxes and testing environments that can be quickly launched and destroyed
- Small business applications: Run business software, blogs, e-commerce sites, and collaboration tools like SharePoint
- Academic and research workloads: Use Lightsail for Research with pre-installed applications like RStudio and Scilab
- Container deployments: Deploy and manage containerized applications using Docker containers
- Email servers: Set up secure email services using software like Postfix, Exim, Dovecot, or Microsoft Exchange

Features

- **Virtual Private Servers (Instances)**: Easy-to-set-up instances backed by AWS reliability with pre-configured OS and applications like WordPress, LAMP, and Windows Server
- **Container Services**: Run and manage containerized applications with automatic scaling and load balancing capabilities
- **Managed Databases**: MySQL, PostgreSQL, and MariaDB databases with automatic backups and high availability options
- **Block and Object Storage**: SSD-based block storage and scalable object storage for additional data storage needs

- **Load Balancers**: Distribute web traffic across instances with session persistence and health checks
- **CDN Distributions**: Content delivery network for caching and delivering content to end-users globally
- **DNS Management**: Domain name system management for registered domains with DNS record configuration
- **Snapshots and Backups**: Manual and automatic snapshots for instances, databases, and block storage for disaster recovery

Value Proposition

Amazon Lightsail offers unmatched simplicity for launching cloud applications with predictable, bundled pricing starting at $0.0067 per hour. Unlike complex AWS services requiring extensive configuration, Lightsail provides pre-configured blueprints that get applications running in minutes:

- The service includes everything needed in one package—compute, storage, networking, and management tools—eliminating the complexity of choosing and configuring multiple AWS services separately
- With built-in features like automatic backups, monitoring, DNS management, and seamless integration with the broader AWS ecosystem through VPC peering, Lightsail provides enterprise-grade capabilities with consumer-friendly simplicity
- The predictable monthly pricing eliminates billing surprises, making it ideal for cost-conscious small businesses and developers who need reliable cloud infrastructure without the operational overhead

Service Integration

Amazon Lightsail integrates seamlessly with core AWS services through VPC peering, enabling access to Amazon EC2, Amazon RDS, Amazon Aurora, AWS Lambda, Amazon S3, Amazon DynamoDB, and Amazon CloudWatch. VPC peering allows Lightsail resources to communicate privately with resources in Amazon VPC, expanding capabilities beyond Lightsail's simplified offerings:

- Instances can be upgraded directly to Amazon EC2 for more advanced requirements
- Lightsail snapshots can be exported to Amazon EC2 for enhanced functionality
- The service works with AWS Identity and Access Management (IAM) for security, AWS CloudTrail for API monitoring, Amazon Route 53 for advanced DNS needs, and integrates with AWS CLI and SDKs for programmatic management
- Container services can connect to Amazon Elastic Container Registry (ECR) for container image storage and management

Onboarding and Offboarding

Getting started with Amazon Lightsail requires completing AWS account setup prerequisites and creating administrative users. Users sign in to the Lightsail console, select an AWS Region and Availability Zone, choose from pre-configured blueprints (Apps + OS like WordPress) or operating systems (OS Only like Ubuntu), select an instance plan based on memory and compute needs, optionally configure networking for dual-stack or IPv6-only, and launch their instance:

- The intuitive management console guides users through configuration with many components pre-configured
- New customers receive 750 hours free on the $5 plan and one free month of managed databases
- Users can connect to instances via browser-based SSH/RDP or traditional clients, add storage, create snapshots, and scale resources as needed through the simplified interface

Getting Started Guide: https://docs.aws.amazon.com/lightsail/latest/userguide/getting-started.html

Data Removal

To offboard from Amazon Lightsail, users must delete all resources including instances, managed databases, load balancers, CDN distributions, and storage volumes. Before deletion, create snapshots of instances and databases to preserve data if needed later. Deleting instances is destructive and removes all data unless snapshots are taken first. Manual snapshots remain in the account after resource deletion and continue incurring storage charges until manually deleted. Automatic snapshots are automatically deleted when source resources are deleted. Users should also remove DNS records, static IP addresses, and container services to avoid ongoing charges.

Backup/Restore and Disaster Recovery

Amazon Lightsail provides comprehensive backup capabilities through manual and automatic snapshots for instances, databases, and block storage disks. Manual snapshots can be created anytime and stored indefinitely, while automatic snapshots are created daily and retain the latest seven copies. Snapshots contain all data needed to restore resources and serve as baselines for creating new resources. System disk snapshots enable file access when instances become unresponsive. Snapshots can be copied across regions within the same AWS account for disaster recovery, and exported to Amazon EC2 for enhanced recovery options.

Backup Capabilities

Amazon Lightsail includes built-in backup capabilities through its snapshot feature but does not directly integrate with AWS Backup service. Lightsail offers both manual and automatic snapshots for instances, managed databases, and block storage disks:

- Automatic snapshots are scheduled daily with seven-day retention, while manual snapshots persist until manually deleted
- Snapshots are incremental, storing only changed blocks for cost efficiency
- The snapshot system operates independently of AWS Backup, providing Lightsail-native backup functionality optimized for the service's simplified management approach

Disaster Recovery

Amazon Lightsail supports disaster recovery through its snapshot and cross-region copy capabilities, but does not integrate with AWS Elastic Disaster Recovery service. Users can create snapshots of instances and databases, copy them to different AWS regions for geographic redundancy, and restore resources from snapshots when needed:

- For more advanced disaster recovery requirements, users can export Lightsail snapshots to Amazon EC2 to leverage AWS Elastic Disaster Recovery and other enterprise-grade recovery services
- The service provides basic disaster recovery suitable for small to medium workloads through its built-in snapshot system

Recovery Objectives

Amazon Lightsail helps customers achieve Recovery Point Objectives (RPO) through daily automatic snapshots, providing up to 24-hour data loss protection, with manual snapshots available for more frequent backup intervals. Recovery Time Objectives (RTO) depend on instance size and data volume, typically ranging from minutes to hours for restoration from snapshots. Cross-region snapshot copying enables geographic redundancy for disaster recovery scenarios. For stricter RPO/RTO requirements, customers can export to Amazon EC2 to access advanced AWS disaster recovery services with sub-minute recovery capabilities and enhanced automation.

Service Constraints

Service Quotas:
https://docs.aws.amazon.com/general/latest/gr/lightsail.html#limits_lightsail

FAQ: https://aws.amazon.com/lightsail/faq/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/lightsail/latest/userguide/what-is-amazon-lightsail.html

User Guide: https://docs.aws.amazon.com/lightsail/latest/userguide/

API Reference: https://docs.aws.amazon.com/lightsail/2016-11-28/api-reference/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/lightsail/pricing/

Cost Optimization

Amazon Lightsail offers unique cost optimization through its predictable bundled pricing model that includes compute, storage, and data transfer allowances in fixed monthly fees. Customers can optimize costs by right-sizing instances using burstable performance that scales automatically based on demand, utilizing the generous data transfer allowances included in each plan, and taking advantage of automatic snapshot retention that manages storage costs by keeping only the latest seven daily backups:

- Container services provide built-in auto-scaling to optimize resource utilization
- The ability to stop instances (releasing public IP) without termination reduces costs during non-use periods
- Free tier offerings include 750 hours of the $5 plan and one month free managed databases for new customers

Savings Considerations

Key cost savings with Amazon Lightsail include avoiding overprovisioning through right-sized instance selection with ability to upgrade via snapshots when needed, leveraging included data transfer allowances that eliminate surprise bandwidth charges common with traditional hosting, utilizing automatic snapshots that prevent expensive data recovery scenarios, and benefiting from simplified management that reduces operational overhead and administrative costs. The bundled pricing model eliminates the complexity and potential cost overruns of configuring multiple AWS services separately:

- Container services with automatic scaling prevent paying for unused capacity
- Block storage starting at $0.10/GB provides cost-effective additional storage
- The predictable monthly billing helps with budget planning and eliminates billing surprises common with usage-based pricing models

Service Name

Amazon Location Service

Service Overview

Amazon Location Service is a fully-managed location data platform that provides developers with geospatial capabilities including maps, places, routes, trackers, and geofences. It enables applications to incorporate location functionality using high-quality data from providers like Esri, HERE, and GrabMaps. The service offers pay-as-you-go pricing with no upfront commitments, provides data privacy and security through anonymization and encryption, and integrates seamlessly with other AWS services for comprehensive location-based solutions.

Key Use Cases

- Asset tracking and fleet management: Monitor vehicles, equipment, and personnel in real-time with geofencing capabilities for operational optimization
- Location-based marketing and customer engagement: Deliver personalized recommendations, proximity-based offers, and location-aware notifications to enhance user experience
- Delivery and logistics optimization: Calculate optimal routes, estimate travel times, and provide real-time tracking for enhanced supply chain visibility and customer satisfaction

Features

- **Maps**: Provides dynamic and static map visualization with multiple pre-designed styles, high-quality base map data, and support for embedding in applications
- **Places**: Offers forward and reverse geocoding, point-of-interest search, address validation, and autocomplete functionality with detailed business information
- **Routes**: Calculates optimal travel routes, estimates travel times, supports multi-point optimization, route matrices, and isoline calculations using real-time traffic data
- **Trackers**: Stores and retrieves current and historical location data for devices with filtering capabilities to reduce storage costs and visual noise
- **Geofences**: Detects when tracked devices enter or exit defined geographical boundaries and triggers automated actions through Amazon EventBridge integration

Value Proposition

Amazon Location Service offers superior data quality from leading providers like Esri, HERE, and GrabMaps, while maintaining full data ownership and privacy through anonymization.

The service provides seamless AWS integration with CloudFormation, CloudWatch, EventBridge, and IAM, enabling faster development and deployment:

- Companies like Halodoc and Geo.me have reduced geospatial costs by 88% and 90% respectively
- The pay-as-you-go model eliminates upfront commitments, and the fully-managed APIs reduce operational overhead while accelerating time-to-market for location-enabled applications

Service Integration

Amazon Location Service integrates natively with core AWS services including Amazon EventBridge for geofence event processing, AWS CloudFormation for infrastructure deployment, Amazon CloudWatch for monitoring and metrics, AWS CloudTrail for audit logging, AWS IAM for access control, and AWS KMS for encryption. The service also works with Amazon DynamoDB for data storage, Amazon S3 for backup storage, and supports AWS SDK integration across multiple programming languages. These integrations enable comprehensive monitoring, security compliance, and automated workflow orchestration.

Onboarding and Offboarding

Customers begin by setting up an AWS account and configuring authentication through API Keys, Amazon Cognito, or AWS IAM. The getting started process involves creating location resources like maps, place indexes, or trackers through the AWS Management Console or APIs:

- Amazon Location provides comprehensive developer tools including standard AWS SDKs, front-end mobile and web SDKs, sample applications, and solution guides
- The interactive Amazon Location Service Console offers visual learning tools, and detailed tutorials guide users through creating their first location-enabled applications using MapLibre GL JS

Getting Started Guide:
https://docs.aws.amazon.com/location/latest/developerguide/getting-started.html

Data Removal

Amazon Location Service automatically deletes tracker position data after 30 days of retention. Geofence geometries must be explicitly deleted by users. When offboarding, customers can delete resources through the AWS Management Console, APIs, or CLI. Deleting an AWS account removes all associated Amazon Location resources and data. The service provides APIs like BatchDeleteDevicePositionHistory for bulk data deletion, ensuring complete data removal during the offboarding process.

Backup/Restore and Disaster Recovery

Amazon Location Service operates as a fully-managed service with built-in high availability across multiple Availability Zones. The service maintains a 99.95% availability SLA with automatic failover capabilities. However, Amazon Location does not provide traditional backup and restore capabilities as it primarily processes real-time location requests rather than storing persistent application data, except for tracker positions which are automatically managed with 30-day retention.

Backup Capabilities

Amazon Location Service does not integrate with AWS Backup as it is primarily a request-response service for location operations. The service automatically manages tracker position data with built-in 30-day retention and deletion policies. For application-specific location data backup needs, customers should implement backup strategies using other AWS services like Amazon S3 or DynamoDB for storing processed location results.

Disaster Recovery

Amazon Location Service does not integrate with AWS Elastic Disaster Recovery as it is a fully-managed, stateless service with built-in multi-AZ redundancy. The service provides automatic failover and high availability through AWS's global infrastructure. For disaster recovery planning, customers should focus on backing up their application configurations, API keys, and any stored location processing results using appropriate AWS storage services.

Recovery Objectives

Amazon Location Service achieves low RPO and RTO through its fully-managed architecture with automatic failover across Availability Zones. The service's 99.95% availability SLA supports sub-minute recovery objectives for location requests. For applications requiring specific RPO/RTO targets, customers should implement caching strategies, configure appropriate retry logic, and design multi-region architectures using Amazon Location's global endpoint availability.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/location/latest/developerguide/manage-quotas.html

FAQ: https://aws.amazon.com/location/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/location/latest/developerguide/index.html

User Guide: https://docs.aws.amazon.com/location/latest/developerguide/what-is.html

API Reference:
https://docs.aws.amazon.com/location/latest/APIReference/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/location/pricing/

Cost Optimization

Amazon Location Service offers multiple cost optimization strategies through tiered pricing buckets (Core, Advanced, Premium) allowing customers to select appropriate feature sets. The service provides filtering options for tracker positions to reduce storage costs by eliminating unnecessary location updates based on distance, time, or accuracy thresholds:

- Customers can optimize Maps API costs by using static maps for non-interactive scenarios, caching map tiles, and selecting appropriate data providers
- The Places API offers Label pricing for basic address text needs, reducing costs for simple geocoding operations

Savings Considerations

Key cost savings opportunities include leveraging the three-month free tier for evaluation and development, using batch APIs like BatchUpdateDevicePosition and BatchGetDevicePosition to reduce per-request costs, implementing local filtering for tracker updates to minimize stored positions, and selecting the most cost-effective pricing bucket for each use case. Volume discounts apply for monthly usage exceeding $5,000. Customers can achieve significant savings by optimizing API request frequency, using appropriate caching strategies, and selecting regional endpoints closest to their user base to reduce latency and associated costs.

Service Name

Amazon Macie

Service Overview

Amazon Macie is a data security service that uses machine learning and pattern matching to discover, monitor, and protect sensitive data in Amazon S3. It automates the discovery and reporting of sensitive data such as personally identifiable information (PII), financial data, and credentials. Macie evaluates S3 buckets for security and access control issues, provides continuous monitoring capabilities, and generates detailed findings for potential security risks. The service integrates with AWS Security Hub and Amazon EventBridge to enable automated remediation workflows and comprehensive security posture management across organizations.

Key Use Cases

- Data privacy compliance: Automatically discover and classify sensitive data to meet regulatory requirements like GDPR, CCPA, and HIPAA
- Security posture monitoring: Continuously evaluate S3 bucket configurations for security and access control vulnerabilities
- Risk assessment and remediation: Identify unencrypted or publicly accessible buckets containing sensitive data for immediate remediation
- Multi-account data governance: Centrally manage sensitive data discovery across organizational accounts using AWS Organizations integration
- Incident response and forensics: Provide detailed location information for sensitive data to support security investigations and audit requirements

Features

- **Automated Sensitive Data Discovery**: Continuously evaluates S3 bucket inventories and analyzes representative objects for sensitive data using machine learning
- **Managed Data Identifiers**: Built-in criteria for detecting over 100 types of sensitive data including PII, financial information, and credentials
- **Custom Data Identifiers**: User-defined regex patterns to detect organization-specific sensitive data types
- **Policy Findings**: Detects changes to S3 bucket policies that reduce security or privacy posture
- **Sensitive Data Discovery Jobs**: On-demand and scheduled analysis of specific S3 buckets with customizable scope and frequency
- **Multi-Account Management**: Centralized management across AWS Organizations with delegated administrator capabilities

- **Allow Lists**: Define text patterns and exceptions that Macie should ignore during analysis
- **Integration Capabilities**: Publishes findings to EventBridge and Security Hub for automated workflows and centralized security management

Value Proposition

Amazon Macie provides unmatched automation and scale for sensitive data discovery in S3 environments, leveraging advanced machine learning to reduce manual effort and false positives. Unlike traditional data loss prevention solutions, Macie offers native AWS integration with seamless setup through a single API call or console selection:

- The service provides comprehensive coverage with over 100 built-in sensitive data types while supporting custom detection patterns for organization-specific requirements
- Macie's multi-account support through AWS Organizations enables centralized governance at enterprise scale, while its pay-as-you-analyze pricing model ensures cost efficiency
- The service's continuous monitoring capabilities and integration with Security Hub and EventBridge create automated security workflows that competitors cannot match in the AWS ecosystem

Service Integration

Amazon Macie integrates deeply with core AWS services to provide comprehensive data security management. It natively monitors Amazon S3 buckets and objects, leveraging S3 bucket policies and access controls for security evaluation:

- Macie publishes findings as events to Amazon EventBridge, enabling automated remediation through Lambda functions, SNS notifications, and Step Functions workflows
- AWS Security Hub integration provides centralized security posture management and compliance reporting across the organization
- AWS Organizations support enables multi-account deployment and management with delegated administrator capabilities
- Macie integrates with AWS CloudTrail for API call logging and audit trails, AWS IAM for access control and permissions management, and AWS KMS for analyzing encrypted objects
- The service also works with AWS Backup for data protection strategies and supports VPC endpoints for private connectivity

Onboarding and Offboarding

Customers can enable Amazon Macie with a single selection in the AWS Management Console or one API call, requiring minimal setup configuration. The service automatically

generates an inventory of S3 general purpose buckets and begins evaluating them for security and access control upon enablement:

- Users must configure IAM permissions for console and API access, then select their desired AWS region and follow on-screen instructions
- Macie immediately starts automated sensitive data discovery, analyzing representative objects in S3 buckets
- For long-term retention of discovery results beyond the default 90-day storage, customers should configure an S3 repository
- Multi-account organizations can enable trusted access between Macie and AWS Organizations, designating a delegated administrator account for centralized management across member accounts

Getting Started Guide: https://docs.aws.amazon.com/macie/latest/user/getting-started.html

Data Removal

When offboarding from Amazon Macie, customers can disable the service through the console or API, which deletes all Macie-specific settings and resources for the account. Sensitive data discovery results are automatically deleted after 90 days unless stored in a customer-managed S3 repository. Customers retain control over findings data published to EventBridge and Security Hub, which must be managed separately. To completely remove data, customers should delete any configured S3 repositories containing discovery results and remove EventBridge rules or Security Hub integrations that process Macie findings.

Backup/Restore and Disaster Recovery

Amazon Macie provides data resiliency through multiple mechanisms but does not offer traditional backup capabilities. The service stores sensitive data discovery results in customer-specified S3 buckets for long-term retention and cross-region replication. Macie publishes findings to EventBridge, allowing customers to send findings data to preferred storage platforms for extended retention. The service leverages AWS global infrastructure with multiple Availability Zones for high availability and fault tolerance. Customers can use Macie API operations to retrieve findings data programmatically and integrate with external backup systems.

Backup Capabilities

Amazon Macie does not provide traditional backup capabilities but offers data retention mechanisms. Discovery results are stored for 90 days by default and can be retained long-term in customer-managed S3 buckets:

- The service does not integrate with AWS Backup as it is not a data storage service but rather analyzes existing S3 data

- Customers must implement their own backup strategies for Macie configurations, custom data identifiers, and findings data through API exports and S3 storage

Disaster Recovery

Amazon Macie does not directly integrate with AWS Elastic Disaster Recovery as it is a security service rather than a workload requiring recovery. The service inherently provides resilience through AWS global infrastructure and multi-region availability. For disaster recovery planning, customers can replicate Macie configurations across regions using Infrastructure as Code, export findings data to multiple regions via S3 cross-region replication, and maintain EventBridge rules in multiple regions for continued security monitoring during regional outages.

Recovery Objectives

Amazon Macie supports recovery objectives by providing automated data discovery that can quickly re-establish security posture visibility after outages. The service's API-driven configuration enables rapid redeployment across regions, supporting low RTO requirements. For RPO objectives, Macie's continuous monitoring and real-time findings publication to EventBridge ensure minimal data loss during incident response scenarios. Organizations can achieve faster recovery by maintaining Macie configurations in multiple regions and implementing cross-region findings replication through S3 and EventBridge integration.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/macie/latest/user/macie-quotas.html

FAQ: https://aws.amazon.com/macie/faq/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/macie/latest/user/what-is-macie.html

User Guide: https://docs.aws.amazon.com/macie/latest/user/getting-started.html

API Reference: https://docs.aws.amazon.com/macie/latest/APIReference/welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/macie/pricing/

Cost Optimization

Amazon Macie offers several unique cost optimization features including automated sensitive data discovery that samples representative objects rather than analyzing entire datasets, reducing analysis costs while maintaining coverage. Customers can exclude specific S3 buckets from automated discovery to focus spending on critical data stores:

- The service provides cost estimation tools during sensitive data discovery job creation, allowing customers to forecast expenses before execution
- Macie's pay-as-you-analyze pricing model with monthly quotas helps control costs, and the 30-day free trial includes up to 150GB of automated discovery analysis
- Organizations can optimize costs by configuring discovery scope, using allow lists to reduce false positives, and leveraging automated discovery over manual jobs where appropriate

Savings Considerations

Customers can achieve significant cost savings with Amazon Macie by leveraging automated discovery instead of running comprehensive manual discovery jobs, which reduces the volume of data analyzed. Strategic use of bucket exclusions and scoped discovery jobs prevents analysis of non-sensitive data repositories:

- The service's machine learning capabilities reduce operational costs by minimizing false positives and manual review efforts
- Macie's integration with existing AWS services eliminates the need for third-party data classification tools and associated licensing costs
- The monthly free tier of 1GB analysis and volume discounts for large organizations provide additional savings opportunities
- Customers should monitor usage through the Macie console's cost estimation features and optimize discovery job frequency based on data change patterns to maximize cost efficiency

Service Name

Amazon Managed Grafana

Service Overview

Amazon Managed Grafana is a fully managed and secure data visualization service that enables users to instantly query, correlate, and visualize operational metrics, logs, and traces from multiple sources. It simplifies the deployment, operation, and scaling of Grafana, a popular open-source data visualization tool. Users create logically isolated Grafana servers called workspaces to build dashboards and visualizations without managing underlying infrastructure. The service handles provisioning, scaling, maintenance, and provides built-in security features including single sign-on, data access control, and audit reporting for corporate governance compliance.

Key Use Cases

- Cloud Infrastructure Monitoring: Monitor AWS resource utilization, cost analysis, and performance metrics across multi-account environments
- Application Performance Management: Create application-specific dashboards for request latency, error rates, and user experience metrics
- Security Operations: Display security events in dedicated dashboards for rapid incident response and investigation
- Operational Observability: Centralize monitoring data from multiple AWS services like CloudWatch, X-Ray, and OpenSearch Service
- DevOps and Team Collaboration: Share interactive dashboards across teams with role-based access control

Features

- **Unified Observability**: Visualize and correlate data across multiple data sources with pre-built panels and dashboards
- **Multi-Account Access**: Secure access to AWS data across multiple accounts and regions using AWS Organizations integration
- **Grafana Alerting**: Robust alerting system with centralized alerting information and multiple alert instances from single rules
- **Team Collaboration**: Easy dashboard sharing with user authentication, authorization, and team-based access control
- **Plugin Management**: Install, update, and manage Grafana community plugins and Enterprise plugins for extended data source support
- **VPC Integration**: Connect to private data sources in VPCs without exposing traffic to the internet

- **Enterprise Security**: Built-in security with SAML 2.0 and AWS IAM Identity Center integration, encryption at rest and in transit

Value Proposition

Amazon Managed Grafana eliminates the operational burden of managing Grafana infrastructure while providing enterprise-grade security and scalability. It offers automatic scaling, high availability, automated version updates, and security patching:

- The service provides superior visualization capabilities compared to CloudWatch with advanced widgets, supports data sources beyond AWS services, and enables separate access management from AWS accounts
- Integration with AWS Organizations allows seamless multi-account data access, while the extensive plugin ecosystem supports diverse data sources including third-party and open-source tools
- The pay-per-active-user pricing model with a 90-day free trial makes it cost-effective for teams of all sizes

Service Integration

Amazon Managed Grafana natively integrates with key AWS observability and data services including Amazon CloudWatch for metrics and logs, Amazon OpenSearch Service for search analytics, AWS X-Ray for distributed tracing, AWS IoT SiteWise for industrial data, Amazon Timestream for time-series data, and Amazon Managed Service for Prometheus for container metrics. It leverages AWS IAM Identity Center and AWS Organizations for authentication and multi-account access management:

- The service uses AWS CloudFormation StackSets for cross-account permission provisioning and integrates with Amazon SNS for alert notifications
- Additional integrations include Amazon Athena for data queries, Amazon RDS for database connections, and AWS VPC for private network access to data sources

Onboarding and Offboarding

Customers begin by creating a workspace in the Amazon Managed Grafana console, selecting authentication methods (AWS IAM Identity Center or SAML), and choosing service-managed permissions for simplified setup. The service automatically provisions necessary IAM roles and policies when AWS data sources are selected:

- Users can enable up to 5 workspaces per region initially
- Setup requires the AWSGrafanaAccountAdministrator policy for workspace creation
- After workspace creation, users sign into the Grafana console to configure data sources using the AWS Data Sources plugin, create dashboards, and manage team access

- The service provides getting started tutorials and pre-built dashboard templates to accelerate initial configuration and time-to-value

Getting Started Guide: https://docs.aws.amazon.com/grafana/latest/userguide/getting-started-with-AMG.html

Data Removal

When deleting an Amazon Managed Grafana workspace, all associated configuration data including dashboards, data source configurations, alerts, and snapshots are permanently removed. Users must navigate to the Amazon Managed Grafana console, select the workspace, and choose Delete with confirmation requiring the workspace name. The service provides API endpoints for programmatic deletion. Note that deleting a workspace does not automatically remove associated IAM roles, which should be cleaned up separately to maintain security best practices.

Backup/Restore and Disaster Recovery

Amazon Managed Grafana provides built-in high availability and automatic recovery capabilities. The service offers snapshot functionality through the Snapshot API for dashboard backups. Snapshots can be stored locally within the workspace or externally on external servers. However, the service does not provide traditional backup and restore capabilities for entire workspaces. Data persistence relies on the managed infrastructure's durability, with dashboards and configurations stored securely in the AWS managed environment with encryption at rest.

Backup Capabilities

Amazon Managed Grafana supports dashboard snapshots through the Snapshot API, allowing users to create point-in-time backups of dashboard configurations and data. Snapshots can be saved locally or externally with configurable expiration times:

- The service does not integrate with AWS Backup for automated backup strategies
- Users can export dashboard configurations using Grafana APIs and Terraform for infrastructure-as-code backup approaches

Disaster Recovery

Amazon Managed Grafana provides high availability within regions through automatic scaling and recovery mechanisms. The service does not integrate with AWS Elastic Disaster Recovery:

- For cross-region disaster recovery, customers must manually recreate workspaces and restore configurations in alternate regions
- The service's managed nature ensures infrastructure-level resilience, but application-level disaster recovery requires manual planning and implementation

Recovery Objectives

Amazon Managed Grafana helps achieve low RTO objectives through its highly available managed infrastructure with automatic failover and recovery. The service provides instant workspace provisioning and rapid dashboard restoration from snapshots. However, RPO objectives depend on manual backup strategies using snapshot APIs or configuration exports. Customers can minimize data loss by implementing regular dashboard exports and maintaining infrastructure-as-code definitions for workspace configurations.

Service Constraints

Service Quotas:
https://docs.aws.amazon.com/grafana/latest/userguide/AMG_quotas.html

FAQ: https://aws.amazon.com/grafana/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/grafana/latest/userguide/what-is-Amazon-Managed-Service-Grafana.html

User Guide: https://docs.aws.amazon.com/grafana/latest/userguide/getting-started-with-AMG.html

API Reference:
https://docs.aws.amazon.com/grafana/latest/APIReference/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/grafana/pricing/

Cost Optimization

Amazon Managed Grafana offers a 90-day free trial with up to 5 free users per account to minimize initial costs. The pay-per-active-user pricing model charges only for users who access the workspace within 24 hours, not provisioned users:

- API requests are free, helping reduce costs for automated workflows
- Customers can optimize costs by using Viewer licenses ($5/month) instead of Editor licenses ($9/month) where appropriate, managing user access patterns to minimize active user counts, and leveraging the free tier for small teams
- Multiple workspaces allow cost segregation between environments

Savings Considerations

Key cost savings include eliminating infrastructure management overhead compared to self-hosted Grafana deployments, reducing operational costs through automatic scaling and maintenance, and avoiding upfront commitments with pay-as-you-use pricing. The service eliminates costs for server provisioning, security patching, version updates, and high availability setup:

- Enterprise plugin licensing provides cost-effective access to premium data sources compared to separate vendor contracts
- Organizations can achieve economies of scale by centralizing monitoring across multiple AWS accounts while maintaining granular access control and cost allocation

Service Name

Amazon Managed Service for Apache Flink

Service Overview

Amazon Managed Service for Apache Flink is a fully managed service that enables users to process and analyze streaming data using Apache Flink, an open-source framework for stream and batch processing. The service handles underlying infrastructure management including resource provisioning, availability zone failover resilience, parallel computation, automatic scaling, and application backups. Users can build applications using Java, Scala, Python, or SQL with support for Apache Flink's DataStream API, Table API, SQL API, and Process Functions. The service offers two deployment options: standard applications for long-running production workloads built with IDEs, and Studio notebooks for interactive real-time data exploration and dashboard creation using Apache Zeppelin notebooks.

Key Use Cases

- Real-time analytics: Processing and analyzing streaming data for immediate insights and time-series analytics
- Event-driven applications: Building responsive applications that react to streaming events with low latency
- Streaming ETL: Transforming and loading streaming data into data warehouses and data lakes in real-time
- Real-time dashboards and metrics: Creating interactive dashboards and monitoring systems with live data updates
- Complex event processing: Detecting patterns and anomalies in streaming data for fraud detection and monitoring
- IoT data processing: Processing high-volume sensor data and telemetry from connected devices
- Clickstream analysis: Real-time analysis of user behavior and website interactions for personalization

Features

- **Multiple Apache Flink APIs**: Support for Flink SQL, Table API, DataStream API, and Process Functions with Java, Scala, Python, and SQL
- **Automatic Scaling**: Auto-scales based on CPU utilization with configurable parallelism and Kinesis Processing Units (KPUs)
- **Fault Tolerance**: Built-in checkpointing, snapshots, and savepoints for application state management and recovery
- **Studio Notebooks**: Interactive Apache Zeppelin notebooks for ad-hoc data exploration and real-time dashboard creation

- **High Availability**: Multi-AZ deployment with automatic failover and resilience across availability zones
- **Per-second Billing**: Pay-per-use pricing model with one-second increments and 10-minute minimum charge
- **Rich Connector Library**: Over 40 pre-built connectors for AWS services and external systems including Kafka, DynamoDB, S3
- **System Rollbacks**: Automatic rollback capabilities for failed updates and scaling operations

Value Proposition

Customers should choose Amazon Managed Service for Apache Flink for its serverless architecture that eliminates infrastructure management overhead while providing enterprise-grade reliability and performance. The service offers significant operational advantages including automatic scaling, built-in fault tolerance through checkpointing and snapshots, and multi-AZ high availability:

- Its per-second billing model with automatic resource optimization reduces costs compared to provisioning fixed infrastructure
- The service provides comprehensive integration with the AWS ecosystem through 40+ pre-built connectors, making it ideal for customers already using AWS services
- Additionally, the dual deployment options - production applications and interactive Studio notebooks - support both development and operational use cases, while the managed nature ensures consistent performance without the complexity of managing Apache Flink clusters

Service Integration

Amazon Managed Service for Apache Flink integrates deeply with core AWS analytics and data services. Primary integrations include Amazon Kinesis Data Streams for real-time data ingestion, Amazon MSK (Managed Streaming for Apache Kafka) for event streaming, Amazon S3 for data storage and batch processing, and Amazon DynamoDB for real-time database operations:

- The service connects to Amazon Data Firehose for data delivery, Amazon OpenSearch Service for search analytics, and Amazon Redshift for data warehousing
- Additional integrations encompass Amazon SQS for messaging, Amazon Managed Service for Prometheus for monitoring, CloudWatch for logging and metrics, AWS Glue for data catalog access, and IAM for security management
- The service also supports AWS KMS for encryption and VPC for network isolation, providing comprehensive integration across the AWS data and analytics ecosystem

Onboarding and Offboarding

Customers get started by first installing prerequisites including Java Development Kit 11, Apache Maven, Git client, and an IDE for development. The process begins with creating an AWS account and setting up the AWS CLI for programmatic access:

- Users then develop Apache Flink applications locally using DataStream API, Table API, or SQL, package them as JAR files, and upload to Amazon S3
- Applications are created through the AWS Console or CLI, configured with runtime properties, and started with specified parallelism settings
- For interactive development, customers can create Studio notebooks directly in the console using Apache Zeppelin for SQL, Python, or Scala development
- The service provides comprehensive getting started tutorials, workshops, and example applications to accelerate onboarding, with step-by-step guides available for different programming languages and use cases

Getting Started Guide: https://docs.aws.amazon.com/managed-flink/latest/java/getting-started.html

Data Removal

Data removal involves stopping applications using the StopApplication API, which automatically creates a final snapshot before termination. Customers must manually delete application snapshots using DeleteApplicationSnapshot API as the service never automatically deletes them. Application code stored in Amazon S3 buckets must be removed separately. VPC configurations, CloudWatch logs, and IAM roles created for the service should be cleaned up manually. The service provides cleanup guidance in documentation to ensure complete data removal and resource decommissioning during offboarding.

Backup/Restore and Disaster Recovery

The service provides comprehensive backup through automatic checkpointing every 60 seconds by default and manual snapshots for long-term state preservation. Checkpoints enable automatic recovery from failures, while snapshots allow point-in-time restoration during updates and scaling operations. Multi-AZ deployment ensures high availability with automatic failover capabilities. System rollback features automatically restore previous application versions during failed operations. All backup data is encrypted and stored in service-managed S3 buckets with configurable retention policies.

Backup Capabilities

Amazon Managed Service for Apache Flink provides built-in backup through checkpointing and snapshots but does not directly integrate with AWS Backup service. The service manages its own backup strategy using Apache Flink's native checkpointing mechanism for automatic recovery and snapshot functionality for manual state preservation:

- Customers can create up to 1,000 snapshots per application for long-term backup retention
- All backup data is automatically encrypted and stored in service-owned S3 buckets, providing durable state preservation without requiring AWS Backup integration

Disaster Recovery

The service provides disaster recovery through multi-AZ high availability deployments and automatic failover mechanisms but does not integrate with AWS Elastic Disaster Recovery service. Managed Service for Apache Flink implements its own disaster recovery strategy using distributed JobManager deployment across availability zones, automatic checkpoint recovery, and snapshot-based restoration. The service's resilient architecture ensures applications can recover from infrastructure failures within the same region through built-in fault tolerance mechanisms rather than requiring external disaster recovery services.

Recovery Objectives

Amazon Managed Service for Apache Flink helps customers achieve low RTO through automatic failover across availability zones and rapid application restart from checkpoints, typically within minutes. RPO objectives are met through configurable checkpoint intervals (default 60 seconds) and manual snapshots for point-in-time recovery. The service's automatic scaling and system rollback features minimize downtime during updates and scaling operations. Multi-AZ deployment and stateful recovery mechanisms ensure minimal data loss and quick service restoration during failures.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/managed-flink/latest/java/limits.html

FAQ: https://aws.amazon.com/managed-service-apache-flink/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/managed-flink/latest/java/what-is.html

User Guide: https://docs.aws.amazon.com/managed-flink/latest/java/getting-started.html

API Reference: https://docs.aws.amazon.com/managed-flink/latest/apiv2/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/managed-service-apache-flink/pricing/

Cost Optimization

Amazon Managed Service for Apache Flink offers several unique cost optimization features including per-second billing with 10-minute minimum charges, eliminating over-provisioning costs. Automatic scaling adjusts KPUs based on actual workload demand, scaling up when CPU utilization exceeds 75% and down when below 10%:

- The service charges only for consumed resources without upfront commitments or infrastructure provisioning costs
- Studio notebooks provide cost-effective development environments for experimentation without dedicated infrastructure
- Customers can optimize costs by right-sizing parallelism settings, using efficient checkpoint intervals, minimizing JAR file sizes, and leveraging automatic scaling
- The serverless model eliminates idle resource costs typical of self-managed Apache Flink clusters, making it particularly cost-effective for variable workloads and development environments

Savings Considerations

Main cost savings come from eliminating infrastructure management overhead and operational complexity of running Apache Flink clusters. The service reduces total cost of ownership by providing built-in high availability, automatic scaling, and managed updates without dedicated DevOps resources:

- Per-second billing ensures customers only pay for actual usage, particularly beneficial for batch workloads and intermittent streaming applications
- Automatic resource optimization prevents over-provisioning while maintaining performance requirements
- The managed service model eliminates costs associated with cluster monitoring, patch management, and capacity planning
- For organizations with variable streaming workloads, the pay-per-use model provides significant savings compared to maintaining always-on infrastructure
- Development and testing costs are reduced through Studio notebooks and the ability to start/stop applications on-demand without infrastructure penalties

Service Name

Amazon Managed Service for Prometheus

Service Overview

Amazon Managed Service for Prometheus is a serverless, Prometheus-compatible monitoring service that makes it easier to securely monitor container environments at scale. It automatically scales the ingestion, storage, and querying of operational metrics as workloads scale up and down, while integrating with AWS security services to enable fast and secure access to data. Built on the open-source Prometheus data model and powered by Cortex, it provides the same PromQL query language and data model used with traditional Prometheus, but without the operational overhead of managing underlying infrastructure. The service is designed for high availability using multi-AZ deployments and replicates data across three Availability Zones.

Key Use Cases

- Container monitoring: Monitor performance and operational metrics from Amazon EKS, Amazon ECS, AWS Fargate, and self-managed Kubernetes clusters at scale
- Application observability: Collect, store, and analyze time-series metrics from containerized applications using familiar Prometheus tools and PromQL queries
- Multi-environment monitoring: Unified monitoring solution across AWS, on-premises, hybrid, and multi-cloud container environments with centralized metric storage and querying

Features

- **Serverless monitoring**: Fully managed service with automatic scaling for ingestion, storage, and querying without infrastructure management
- **Prometheus compatibility**: Native support for Prometheus data model, PromQL query language, and over 150+ Prometheus exporters
- **AWS managed collectors**: Agentless metric collection that automatically discovers and pulls metrics from Kubernetes clusters through ENIs
- **Alert Manager**: Built-in alerting with deduplication, grouping, routing to SNS topics, and integration with PagerDuty
- **High availability**: Multi-AZ deployments with data replication across three Availability Zones and automatic failover
- **Security integration**: Encryption in transit and at rest, IAM access control, AWS KMS integration, and VPC endpoint support
- **Query insights and control**: Query logging to CloudWatch, cost control with QSP thresholds, and expensive query monitoring

Value Proposition

Amazon Managed Service for Prometheus eliminates the operational overhead of managing Prometheus infrastructure while providing enterprise-grade scalability, security, and availability. Customers benefit from automatic scaling that handles up to 1 billion active series per workspace, native integration with AWS security services, and seamless compatibility with existing Prometheus tooling:

- The service offers significant cost advantages through efficient resource utilization and pay-as-you-go pricing, while enabling teams to focus on application development rather than monitoring infrastructure maintenance
- Integration with Amazon Managed Grafana provides a complete managed observability solution, and the service's multi-AZ architecture ensures high availability without additional configuration effort

Service Integration

Amazon Managed Service for Prometheus deeply integrates with Amazon EKS and Amazon ECS for container metrics collection, Amazon Managed Grafana for visualization and dashboards, and AWS Distro for OpenTelemetry (ADOT) for standardized metric collection. It leverages Amazon CloudWatch for usage metrics monitoring and alerting, Amazon SNS for alert notifications, AWS IAM for access control, and AWS KMS for encryption key management:

- The service supports VPC endpoints for private network access and integrates with AWS CloudTrail for API call logging
- Additional integrations include Amazon MSK for Kafka monitoring, AWS Lambda for custom processing, and AWS Cost and Usage Reports for billing analysis

Onboarding and Offboarding

Getting started requires three key steps: creating a workspace to store metrics, ingesting metrics either manually or through automatic scraping, and querying the data using PromQL. Customers can create workspaces through the AWS Management Console, AWS CLI, or SDKs, then configure metric ingestion using AWS managed collectors for EKS clusters or custom Prometheus agents:

- The service supports both agentless collection through managed scrapers and traditional Prometheus remote write configurations
- Integration with Amazon Managed Grafana provides immediate visualization capabilities, while alert rules can be configured for proactive monitoring
- AWS provides comprehensive documentation, getting started guides, and example configurations to streamline the onboarding process

Getting Started Guide: https://docs.aws.amazon.com/prometheus/latest/userguide/AMP-getting-started.html

Data Removal

Data removal occurs through workspace deletion, which permanently removes all stored metrics and metadata. Customers must delete any active managed scrapers before deleting workspaces to prevent continued metric ingestion and charges. Data is automatically deleted after the retention period (150 days by default, configurable up to 1095 days). The service provides configurable retention policies per workspace, and customers can monitor usage through CloudWatch metrics to ensure complete data removal after offboarding.

Backup/Restore and Disaster Recovery

Amazon Managed Service for Prometheus provides built-in high availability through multi-AZ deployments with automatic data replication across three Availability Zones. The service includes automatic patch management and infrastructure maintenance. For disaster recovery, customers can implement cross-region data replication by configuring multiple workspaces in different regions and using Prometheus remote write to send metrics to secondary workspaces, ensuring data availability even when primary regions are unavailable.

Backup Capabilities

The service does not integrate directly with AWS Backup but provides data resilience through automatic replication across multiple Availability Zones within a region. Data durability is ensured through the underlying AWS infrastructure, and customers can implement backup strategies by replicating metrics to secondary workspaces in different regions or using Prometheus remote write configurations to maintain copies of critical metrics data.

Disaster Recovery

Amazon Managed Service for Prometheus does not integrate with AWS Elastic Disaster Recovery but provides built-in disaster recovery capabilities through multi-AZ architecture and cross-region workspace replication. The service automatically handles infrastructure failures and provides high availability within regions. For comprehensive disaster recovery, customers can configure secondary workspaces in different AWS regions and implement cross-region metric replication using Prometheus remote write functionality.

Recovery Objectives

The service helps meet RPO and RTO objectives through automatic multi-AZ replication providing near-zero RPO within regions, and sub-minute RTO for automatic failover scenarios. Cross-region workspace configurations can achieve broader disaster recovery objectives with RPO dependent on metric ingestion frequency and RTO based on DNS failover or application-level switching between regions. The serverless architecture enables rapid recovery without infrastructure provisioning delays.

Service Constraints

Service Quotas:
https://docs.aws.amazon.com/prometheus/latest/userguide/AMP_quotas.html

FAQ: https://aws.amazon.com/prometheus/faqs/

Technical Requirements

Service Documentation:
https://docs.aws.amazon.com/prometheus/latest/userguide/what-is-Amazon-Managed-Service-Prometheus.html

User Guide: https://docs.aws.amazon.com/prometheus/latest/userguide/

API Reference: https://docs.aws.amazon.com/prometheus/latest/APIReference/

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/prometheus/pricing/

Cost Optimization

Amazon Managed Service for Prometheus offers several unique cost optimization features including label-based active series limits to control metric volume by producer, query insights and control with QSP thresholds to manage expensive queries, and tiered pricing that scales with usage. The service provides CloudWatch usage metrics for monitoring ingestion rates and active series, enabling proactive cost management:

- Customers can optimize costs by reducing collection frequency, limiting active series through metric filtering, using native alerting instead of external systems, and implementing query optimization techniques
- The pay-as-you-go model eliminates infrastructure overhead costs while automatic scaling prevents over-provisioning

Savings Considerations

Primary cost savings come from eliminating self-managed Prometheus infrastructure, reducing operational overhead, and leveraging automatic scaling to match actual usage. The service's tiered pricing model provides cost efficiency at scale, while managed collectors reduce agent maintenance costs:

- Organizations can achieve significant savings through reduced engineering time for monitoring infrastructure, elimination of storage and compute costs for self-hosted solutions, and optimized query performance

- AWS Free Tier provides 40 million samples for ingestion, storage, and querying monthly, enabling cost-effective evaluation and small-scale deployments
- Cross-region replication costs should be considered for disaster recovery implementations, balanced against the operational savings from managed infrastructure

Service Name

Amazon MSK

Service Overview

Amazon Managed Streaming for Apache Kafka (Amazon MSK) is a fully managed service that makes it easy to build and run applications using Apache Kafka to process streaming data. MSK eliminates the operational overhead of managing Apache Kafka infrastructure by automatically creating and managing broker and ZooKeeper nodes, enabling developers to focus on building applications. The service is compatible with open-source Apache Kafka versions and supports existing applications, tooling, and plugins. MSK provides control-plane operations for cluster management while allowing users to leverage Apache Kafka data-plane operations for producing and consuming data. It offers both provisioned and serverless deployment options to meet varying workload requirements.

Key Use Cases

- Real-time data streaming and processing: Capturing and processing application log files, website clickstreams, and IoT sensor data in real-time
- Event-driven microservices architecture: Building decoupled, scalable applications where services communicate through event streams
- Data ingestion for analytics: Streaming data to data lakes and warehouses for real-time analytics and machine learning
- Financial transaction processing: Tracking and processing financial transactions with low latency requirements
- Content delivery and media streaming: Distributing real-time content updates and media streams to applications

Features

- **MSK Provisioned**: User-controlled broker scaling with Standard and Express broker options for optimized performance and cost
- **MSK Serverless**: Fully managed, auto-scaling Kafka clusters without capacity planning or infrastructure management
- **Multi-AZ Deployment**: High availability with automatic failover and data replication across multiple Availability Zones
- **MSK Connect**: Fully managed Kafka Connect service for streaming data between Kafka and external systems
- **MSK Replicator**: Cross-region and cross-cluster data replication for disaster recovery and multi-region architectures
- **Enterprise Security**: Encryption at rest and in transit, IAM authentication, VPC isolation, and granular access control

- **Monitoring and Observability**: CloudWatch metrics, JMX metrics export to Prometheus, and broker log streaming
- **Tiered Storage**: Cost-optimized storage with automatic data tiering to reduce storage costs while maintaining performance
- **Auto Scaling**: Automatic broker and storage scaling based on demand for both provisioned and serverless clusters

Value Proposition

Amazon MSK eliminates the undifferentiated heavy lifting of managing Apache Kafka infrastructure, allowing organizations to focus on application development rather than cluster operations. MSK provides superior operational efficiency by handling patching, scaling, security, and failure recovery automatically, reducing management overhead by up to 80% compared to self-managed solutions:

- The service offers flexible deployment options with both provisioned and serverless models to optimize costs based on usage patterns
- MSK ensures enterprise-grade security with built-in encryption, IAM integration, and VPC isolation without additional configuration complexity
- Deep AWS service integrations enable seamless data pipelines with Lambda, EMR, Flink, S3, and other services, accelerating time-to-market for streaming applications

Service Integration

Amazon MSK integrates deeply with the AWS ecosystem to enable comprehensive data streaming architectures. Core integrations include AWS Lambda for serverless event processing with automatic scaling and built-in event source mappings, Amazon EMR and Amazon Managed Service for Apache Flink for large-scale stream processing and analytics, Amazon S3 for data lake storage through Firehose integration, and AWS Glue for ETL operations on streaming data:

- Additional integrations encompass Amazon CloudWatch for monitoring and alerting, Amazon VPC for secure network isolation, AWS IAM for authentication and authorization, Amazon EventBridge Pipes for event routing, and Amazon Athena for querying processed data
- These integrations eliminate the need for complex custom connectors and provide a unified, managed experience across the entire data streaming pipeline

Onboarding and Offboarding

Getting started with Amazon MSK involves a straightforward process through the AWS Management Console, CLI, or APIs. Customers begin by creating a VPC and subnets, then provision an MSK cluster by selecting either provisioned or serverless deployment options:

- The service provides step-by-step tutorials including cluster creation, IAM role configuration for topic access, client machine setup, and topic creation
- MSK supports standard Apache Kafka clients and tools, enabling existing applications to connect with minimal changes
- The getting started guide includes producing and consuming data samples, monitoring setup through CloudWatch, and cleanup procedures
- AWS provides comprehensive documentation, hands-on labs, and migration guides to accelerate adoption
- For organizations with existing Kafka deployments, MSK offers migration tools and best practices to transition workloads seamlessly

Getting Started Guide: https://docs.aws.amazon.com/msk/latest/developerguide/getting-started.html

Data Removal

Data removal from Amazon MSK requires deleting topics, consumer groups, and cluster resources through the Kafka APIs or AWS Management Console. When offboarding, customers should first stop all producers and consumers, then delete topics containing sensitive data using Kafka administrative commands. The MSK cluster itself can be deleted through the AWS Console, CLI, or API, which removes all associated broker nodes, ZooKeeper instances, and cluster metadata. Storage volumes are automatically deleted when the cluster is terminated. For compliance requirements, customers should ensure proper data retention policies are configured and verify complete data deletion through AWS CloudTrail logs.

Backup/Restore and Disaster Recovery

Amazon MSK provides disaster recovery capabilities primarily through MSK Replicator for cross-region replication and built-in multi-AZ deployment for regional resilience. The service automatically replicates data across Availability Zones within a region for high availability. For cross-region disaster recovery, MSK Replicator continuously replicates topics to secondary regions with configurable replication settings. MSK integrates with AWS Backup for operational backup scenarios, though Kafka's distributed nature makes traditional backup less critical. Recovery objectives are supported through automated failover mechanisms and the ability to quickly provision replacement clusters from replicated data in alternate regions.

Backup Capabilities

Amazon MSK does not integrate directly with AWS Backup service due to Kafka's distributed streaming architecture where data durability comes from replication rather than traditional backups. Instead, MSK provides data protection through automatic multi-AZ replication within clusters, configurable topic-level replication factors, and MSK Replicator for cross-region data copying. The service focuses on continuous data

availability and streaming resilience rather than point-in-time backup recovery, which aligns with Apache Kafka's design principles for handling streaming data workloads.

Disaster Recovery

Amazon MSK does not directly integrate with AWS Elastic Disaster Recovery (DRS) as it is designed for streaming data rather than traditional disaster recovery scenarios. Instead, MSK provides disaster recovery through MSK Replicator for cross-region replication, multi-AZ deployment for zone-level resilience, and automatic cluster recovery capabilities:

- Organizations implement disaster recovery by deploying MSK clusters in multiple regions, using MSK Replicator to maintain data synchronization, and implementing application-level failover logic to redirect traffic during regional impairments
- This approach is specifically designed for streaming workloads that require continuous data flow rather than point-in-time recovery

Recovery Objectives

Amazon MSK helps customers meet Recovery Point Objective (RPO) and Recovery Time Objective (RTO) requirements through multiple mechanisms. For RPO, MSK provides configurable replication factors and MSK Replicator for cross-region replication with minimal data loss potential. Multi-AZ deployment ensures sub-second RPO within regions through synchronous replication. For RTO, MSK offers automatic broker replacement and fast cluster recovery typically within minutes. MSK Replicator enables quick failover to secondary regions, and the serverless option provides instant scaling without capacity planning delays, supporting aggressive RTO targets for business-critical streaming applications.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/msk/latest/developerguide/limits.html

FAQ: https://aws.amazon.com/msk/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/msk/latest/developerguide/what-is-msk.html

User Guide: https://docs.aws.amazon.com/msk/latest/developerguide/getting-started.html

API Reference: https://docs.aws.amazon.com/msk/1.0/apireference/

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/msk/pricing/

Cost Optimization

Amazon MSK offers unique cost optimization through Express brokers that provide higher throughput and faster scaling at optimized costs compared to Standard brokers. MSK Serverless eliminates idle resource costs by charging only for actual usage without minimum capacity requirements:

- Tiered storage automatically moves older data to cost-effective storage tiers while maintaining access performance
- The service provides granular pricing models including separate charges for compute, storage, and data transfer, enabling fine-tuned cost control
- MSK Connect and MSK Replicator offer usage-based pricing for data integration and replication needs
- Organizations can optimize costs by selecting appropriate broker types, configuring retention policies, and leveraging auto-scaling capabilities to match capacity with demand patterns

Savings Considerations

Primary cost savings strategies for Amazon MSK include selecting MSK Serverless for variable workloads to eliminate over-provisioning costs and using Express brokers for predictable workloads requiring higher performance. Implementing proper data retention policies and leveraging tiered storage can significantly reduce storage costs over time:

- Organizations should optimize partition counts and replication factors based on actual durability requirements rather than using maximum settings
- Cross-region data transfer costs can be managed by strategic placement of clusters and careful planning of MSK Replicator usage
- Regular monitoring of CloudWatch metrics helps identify underutilized resources and opportunities for downsizing
- Consolidating multiple small clusters into larger, more efficient deployments can reduce per-broker overhead costs while maintaining isolation through topic-level access controls

Service Name

Amazon Managed Workflows for Apache Airflow

Service Overview

Amazon Managed Workflows for Apache Airflow (MWAA) is a fully managed orchestration service for Apache Airflow that makes it easier to set up and operate end-to-end data pipelines in the cloud at scale. It simplifies managing Apache Airflow by automatically setting up and scaling infrastructure, providing built-in authentication and security, and integrating with various AWS services. Amazon MWAA uses Apache Airflow and Python to create workflows without managing underlying infrastructure, automatically scaling workers based on workload demands while maintaining the same open-source Apache Airflow user interface and functionality.

Key Use Cases

- Batch data processing workflows: Processing large datasets in scheduled batches with complex dependencies and orchestration requirements
- ETL pipeline orchestration: Coordinating extract, transform, and load operations across multiple data sources and destinations
- Multi-service workflow coordination: Orchestrating tasks across various AWS services like S3, EMR, Glue, Lambda, and SageMaker for comprehensive data processing
- Cross-account and cross-region pipeline management: Building unified pipelines spanning multiple AWS accounts and regions while maintaining compliance
- Machine learning workflow orchestration: Managing ML training pipelines, model deployment, and inference workflows with data dependencies
- Legacy workflow migration: Transitioning from AWS Data Pipeline or self-managed Airflow deployments to managed infrastructure

Features

- **Automatic Airflow Setup**: Quickly deploys Apache Airflow environments with chosen versions using the same open-source UI and code
- **Automatic Scaling**: Automatically scales workers and webservers based on workload demands with configurable minimum and maximum limits
- **Built-in Security**: Runs in Amazon VPC with automatic data encryption using AWS KMS, IAM integration, and role-based authentication
- **Public/Private Access Modes**: Supports both public internet access and private VPC-only access to Apache Airflow webserver
- **Streamlined Upgrades**: Provides automatic patches and managed upgrades to new Apache Airflow versions

- **Workflow Monitoring**: Integrates with CloudWatch for logs, metrics, and monitoring without third-party tools
- **AWS Service Integration**: Native integration with 20+ AWS services including S3, EMR, Glue, Lambda, and hundreds of operators
- **Serverless Option**: MWAA Serverless provides pay-per-use pricing with automatic infrastructure scaling

Value Proposition

Amazon MWAA eliminates the operational overhead of managing Apache Airflow infrastructure while providing enterprise-grade security, scalability, and reliability. Customers benefit from reduced operational costs by removing the need to manage open-source Apache Airflow at scale, automatic scaling that handles variable workloads efficiently, and seamless integration with AWS services:

- The managed service provides built-in security with VPC isolation and KMS encryption, streamlined upgrades without maintenance windows, and comprehensive monitoring through CloudWatch
- MWAA maintains full compatibility with Apache Airflow APIs and supports both traditional environments and serverless deployments, offering flexibility in pricing models and deployment options that self-managed solutions cannot match

Service Integration

Amazon MWAA integrates natively with over 20 core AWS services through built-in operators and connectors. Primary integrations include Amazon S3 for DAG storage and data operations, Amazon EMR for big data processing, AWS Glue for ETL workflows, AWS Lambda for serverless task execution, and Amazon Redshift for data warehousing:

- Additional integrations span Amazon Athena for analytics, AWS Batch for batch computing, Amazon DynamoDB for NoSQL operations, Amazon EKS and ECS for container orchestration, Amazon SageMaker for machine learning workflows, and messaging services like SQS and SNS
- The service also integrates with AWS security and monitoring services including IAM, KMS, CloudWatch, and CloudTrail, enabling comprehensive workflow orchestration across the entire AWS ecosystem

Onboarding and Offboarding

Getting started with Amazon MWAA requires creating an Amazon S3 bucket with versioning enabled for DAG files, configuring an Amazon VPC with appropriate subnets and security groups, and creating the MWAA environment through the AWS console, CLI, or CloudFormation. Customers specify the Apache Airflow version, S3 bucket location for DAGs, VPC networking configuration, and optional settings for plugins, requirements, and security:

- The environment takes approximately 20-30 minutes to create and automatically provisions schedulers, workers, and webservers as Fargate containers
- Prerequisites include AWS account permissions, VPC setup, and S3 bucket configuration
- Multiple deployment options are available including quick-start CloudFormation templates and incremental setup guides for different VPC networking scenarios

Getting Started Guide: https://docs.aws.amazon.com/mwaa/latest/userguide/get-started.html

Data Removal

When offboarding from Amazon MWAA, customers can delete their environment through the AWS console, CLI, or API, which automatically removes all managed infrastructure including schedulers, workers, and webservers. The metadata database and associated VPC endpoints are also cleaned up. However, customers retain full control over their DAG files, plugins, and requirements stored in S3, which must be manually deleted if desired. CloudWatch logs and metrics can be retained or deleted based on configured retention policies, ensuring customers maintain control over their workflow history and operational data.

Backup/Restore and Disaster Recovery

Amazon MWAA provides built-in backup capabilities through its managed metadata database with automatic snapshots and point-in-time recovery. The service stores DAGs and supporting files in ma on S3, hich provides ’s of durability and cross-region replication capabilities. hile doesn’t directly integrate ith S ac up service, customers can implement disaster recovery by replicating S3 buckets across regions and recreating environments. The managed nature ensures automatic patching and infrastructure redundancy within regions, while customers can design cross-region DR strategies using S3 replication and Infrastructure as Code approaches.

Backup Capabilities

Amazon MWAA includes automatic backup capabilities through its managed metadata database with built-in snapshots and point-in-time recovery. DAG files and supporting artifacts are stored in Amazon S3, providing inherent durability and versioning capabilities:

- However, MWAA does not directly integrate with AWS Backup service for centralized backup management
- Customers can implement backup strategies using S3 versioning, cross-region replication, and Infrastructure as Code templates to recreate environments
- The service automatically maintains environment configuration and can be restored through CloudFormation or Terraform templates

Disaster Recovery

Amazon MWAA does not directly integrate with AWS Elastic Disaster Recovery service. However, the service supports disaster recovery through its architecture using S3 for DAG storage with cross-region replication capabilities and managed metadata database with automatic backups:

- Customers can implement disaster recovery by replicating S3 buckets to secondary regions and using Infrastructure as Code to quickly recreate MWAA environments
- The containerized architecture on Fargate provides built-in fault tolerance within regions, while multi-region strategies require customer-implemented cross-region replication and environment recreation processes

Recovery Objectives

Amazon MWAA helps customers meet RPO and RTO objectives through its managed architecture and S3-based storage model. RPO can be minimized to near-zero using S3 cross-region replication for DAG files and continuous metadata database backups. RTO depends on environment recreation time (20-30 minutes) plus S3 replication lag. The service supports backup and restore DR strategies with RTOs of 2-4 hours, suitable for most workflow orchestration requirements. Customers requiring faster recovery can implement warm standby architectures using multiple regions with pre-created environments and automated failover mechanisms.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/mwaa/latest/userguide/mwaa-quotas.html

FAQ: https://aws.amazon.com/managed-workflows-for-apache-airflow/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/mwaa/latest/userguide/what-is-mwaa.html

User Guide: https://docs.aws.amazon.com/mwaa/latest/userguide/get-started.html

API Reference: https://docs.aws.amazon.com/mwaa/latest/API/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/managed-workflows-for-apache-airflow/pricing/

Cost Optimization

Amazon MWAA offers unique cost optimization through MWAA Serverless with pay-per-use pricing based on actual task execution time, eliminating costs for idle infrastructure. The traditional MWAA model supports automatic scaling of workers and webservers to match workload demands, preventing over-provisioning:

- Customers can optimize costs by choosing appropriate environment classes (micro, small, medium, large), configuring minimum/maximum worker counts, and utilizing the micro environment for development and testing
- The service eliminates operational overhead costs of managing self-hosted Airflow infrastructure including patching, monitoring, and scaling
- Cost control is enhanced through precise resource allocation and the ability to schedule maintenance windows to minimize disruption

Savings Considerations

Key cost savings opportunities include using MWAA Serverless for sporadic workloads to avoid paying for idle time, selecting the smallest appropriate environment class like mw1.micro for development and light production workloads, and implementing proper worker autoscaling configurations to minimize over-provisioning. Customers save significantly on operational costs by eliminating the need for dedicated infrastructure management teams, automated patching, and 24/7 monitoring:

- Right-sizing worker counts based on actual workload patterns, using efficient DAG designs to reduce task execution time, and leveraging AWS service integrations instead of external tools contribute to cost optimization
- The managed service model eliminates capital expenses for infrastructure while providing predictable operational expenses based on actual usage patterns

Service Name

Amazon MQ

Service Overview

Amazon MQ is a managed message broker service for Apache ActiveMQ Classic and RabbitMQ that enables software applications to communicate using various programming languages, operating systems, and messaging protocols. The service eliminates the operational burden of managing message broker infrastructure by automating setup, maintenance, security patches, and version upgrades. Amazon MQ supports industry-standard APIs and protocols including JMS, NMS, AMQP, STOMP, MQTT, and WebSocket, allowing applications to migrate to the cloud without rewriting messaging code. Organizations can access ActiveMQ and RabbitMQ web consoles directly while benefiting from AWS-managed infrastructure, automated failover, and integration with other AWS services.

Key Use Cases

- Enterprise application modernization: Migrate existing on-premises message brokers to the cloud without rewriting messaging code
- Microservices decoupling: Enable loose coupling between application components using reliable message queuing and pub/sub patterns
- Event-driven architectures: Process real-time events and data streams through message routing and transformation workflows

Features

- **Managed Infrastructure**: Automated provisioning, patching, maintenance, and scaling of message broker infrastructure
- **Multi-Engine Support**: Support for Apache ActiveMQ Classic and RabbitMQ with native web consoles and APIs
- **High Availability**: Active/standby deployments for ActiveMQ and cluster deployments for RabbitMQ across multiple AZs
- **Industry Standard Protocols**: Support for JMS, NMS, AMQP 0.9.1, AMQP 1.0, STOMP, MQTT, and WebSocket protocols
- **Security and Encryption**: VPC access, TLS/SSL encryption, message encryption at rest and in transit with AWS KMS
- **Storage Options**: Durability-optimized with Amazon EFS and throughput-optimized with Amazon EBS storage
- **Network of Brokers**: ActiveMQ broker mesh topology for horizontal scaling and message routing

- **Monitoring Integration**: CloudWatch metrics, logs, and alarms for broker performance monitoring

Value Proposition

Amazon MQ eliminates the undifferentiated heavy lifting of managing message broker infrastructure while providing full compatibility with existing messaging applications. Customers benefit from automated maintenance, security patching, and version upgrades without operational overhead:

- The service offers high availability through multi-AZ deployments, integrated security with AWS services, and cost optimization through managed storage options
- Unlike self-managed solutions, Amazon MQ includes inter-node data transfer at no additional cost for RabbitMQ and provides seamless integration with AWS services like Lambda, CloudWatch, and VPC
- Organizations can migrate existing ActiveMQ or RabbitMQ applications without code changes while gaining enterprise-grade reliability and AWS security compliance

Service Integration

Amazon MQ integrates natively with AWS Lambda for event-driven processing, allowing functions to consume messages from brokers automatically. CloudWatch provides comprehensive monitoring through metrics, logs, and alarms for broker performance tracking:

- Amazon VPC enables secure private network access with security groups and subnet configurations
- AWS Key Management Service (KMS) provides encryption key management for message encryption at rest and in transit
- Amazon EFS and EBS provide managed storage options for message persistence
- EventBridge Pipes can consume messages from Amazon MQ brokers for event routing
- AWS PrivateLink enables private API connectivity without internet exposure
- IAM provides fine-grained access control and authentication policies

Onboarding and Offboarding

Customers begin by accessing the Amazon MQ console and selecting either Apache ActiveMQ or RabbitMQ as the broker engine. The setup process involves choosing deployment mode (single-instance, active/standby for ActiveMQ, or cluster for RabbitMQ), configuring broker instance types, setting up storage options, and creating broker credentials:

- Network configuration includes VPC settings, security groups, and subnet assignments
- Applications connect by updating endpoint URLs to point to the Amazon MQ broker endpoints
- The service provides wire-level protocol endpoints for different messaging protocols
- Most existing applications require only endpoint configuration changes without code modifications
- Getting started tutorials and migration guides are available to assist with the transition process

Getting Started Guide: https://docs.aws.amazon.com/amazon-mq/latest/developer-guide/getting-started-activemq.html

Data Removal

Customers can delete Amazon MQ brokers through the AWS Management Console, API, or CLI by selecting the broker and confirming deletion. The deletion process takes approximately 5 minutes and permanently removes the broker and associated data. For cross-region data replication (CRDR) brokers, customers must first unpair and reboot both primary and replica brokers before deletion. All message data, configurations, and broker metadata are permanently removed during deletion. Customers should ensure proper data backup or migration before initiating deletion as the process is irreversible.

Backup/Restore and Disaster Recovery

Amazon MQ provides data durability through storage replication across multiple Availability Zones using Amazon EFS for durability-optimized storage and Amazon EBS snapshots for throughput-optimized storage. Active/standby brokers for ActiveMQ and cluster deployments for RabbitMQ provide high availability and automatic failover capabilities. Cross-region data replication (CRDR) enables data replication between brokers in different AWS regions for disaster recovery scenarios. Message persistence is maintained through the underlying storage systems with built-in redundancy.

Backup Capabilities

Amazon MQ does not directly integrate with AWS Backup service. Instead, the service provides data durability through managed storage solutions including Amazon EFS replication across multiple AZs for durability-optimized brokers and Amazon EBS snapshots for throughput-optimized brokers:

- Message persistence and broker configurations are maintained through the underlying AWS storage infrastructure
- Customers can implement custom backup strategies for broker configurations and message data using Amazon MQ APIs and storage-level backup mechanisms

Disaster Recovery

Amazon MQ does not directly integrate with AWS Elastic Disaster Recovery service. The service provides disaster recovery capabilities through cross-region data replication (CRDR) for ActiveMQ brokers, enabling real-time data replication between primary and replica brokers in different regions:

- Multi-AZ deployments with automatic failover provide regional disaster recovery
- RabbitMQ cluster deployments distribute nodes across multiple Availability Zones for high availability within a region
- Customers can implement multi-region architectures using network of brokers or application-level replication strategies

Recovery Objectives

Amazon MQ supports various RPO and RTO objectives through deployment configurations. Active/standby brokers provide near-zero RPO with automatic failover achieving RTO of minutes. Cross-region data replication (CRDR) enables regional disaster recovery with minimal data loss. Single-instance brokers offer basic availability with restart capabilities. Cluster deployments for RabbitMQ provide node-level failover with shared state across multiple nodes. Storage durability through Amazon EFS and EBS provides data persistence and recovery capabilities.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/amazon-mq/latest/developer-guide/amazon-mq-limits.html

FAQ: https://aws.amazon.com/amazon-mq/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/amazon-mq/latest/developer-guide/welcome.html

User Guide: https://docs.aws.amazon.com/amazon-mq/latest/developer-guide/welcome.html

API Reference: https://docs.aws.amazon.com/amazon-mq/latest/api-reference/resources.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/amazon-mq/pricing/

Cost Optimization

Amazon MQ offers several cost optimization strategies including rightsizing broker instances based on actual message throughput and connection requirements. Customers can choose between durability-optimized storage using Amazon EFS for high availability needs or throughput-optimized storage using Amazon EBS for performance-focused workloads:

- The service eliminates inter-node data transfer charges for RabbitMQ clusters, providing significant cost savings compared to self-managed deployments
- Single-instance brokers offer lower costs for development and testing environments
- Storage costs can be optimized by configuring appropriate message retention policies and monitoring actual storage utilization
- The pay-as-you-use pricing model eliminates upfront infrastructure costs and provides cost predictability

Savings Considerations

Primary cost savings come from eliminating the operational overhead of managing message broker infrastructure including patching, monitoring, and maintenance activities. The fully managed service reduces staffing requirements for infrastructure management and reduces the time-to-market for messaging applications:

- Inter-node data transfer is included at no additional cost for RabbitMQ, providing significant savings for clustered deployments compared to self-managed solutions
- Storage optimization through managed EFS and EBS reduces storage management complexity and costs
- The service eliminates the need for dedicated infrastructure for broker management and monitoring
- AWS Free Tier provides up to $200 in credits for new customers, enabling cost-effective evaluation and small-scale deployments

Service Name

Amazon Neptune

Service Overview

Amazon Neptune is a fast, reliable, fully managed graph database service that makes it easy to build and run applications that work with highly connected datasets. The core of Neptune is a purpose-built, high-performance graph database engine optimized for storing billions of relationships and querying the graph with milliseconds latency. Neptune supports popular graph query languages including Apache TinkerPop Gremlin, Neo4j's openCypher, and W3C's RDF query language SPARQL. It is highly available with read replicas, point-in-time recovery, continuous backup to Amazon S3, and replication across Availability Zones. Neptune eliminates database management tasks like hardware provisioning, software patching, setup, configuration, and backups.

Key Use Cases

- Recommendation engines: Building personalized content and product recommendation systems using graph relationships
- Fraud detection: Identifying fraudulent patterns and activities through complex relationship analysis in real-time
- Knowledge graphs: Creating structured repositories of information with defined concepts, properties, and relationships
- Identity graphs: Building unified customer views across multiple devices and identifiers for personalization and advertising
- Network security: Analyzing network relationships and dependencies to detect threats and vulnerabilities
- Drug discovery: Modeling molecular relationships and interactions for pharmaceutical research
- Social networking: Managing complex social relationships and connections between users

Features

- **Multi-model Graph Support**: Supports both Property Graph and W3C RDF graph models with query languages Gremlin, openCypher, and SPARQL
- **High Performance**: Optimized for storing billions of relationships with millisecond latency queries and up to 100,000 queries per second
- **Neptune Serverless**: Automatically provisions and scales database capacity based on workload requirements with pay-per-use pricing
- **Global Database**: Spans multiple AWS Regions with storage-based replication, sub-second latency, and disaster recovery capabilities

- **Multi-AZ Deployment**: Provides high availability with read replicas across Availability Zones and automatic failover
- **Neptune ML**: Integrates machine learning using graph neural networks for node classification, edge prediction, and link prediction
- **Continuous Backup**: Automatic, continuous, incremental backups with point-in-time recovery and database snapshots
- **Advanced Security**: Network isolation with VPC, encryption at rest and in transit, IAM integration, and compliance certifications
- **Neptune Analytics**: Memory-optimized analytics engine for analyzing large graph datasets with built-in algorithms
- **GraphRAG**: Combines graph databases with generative AI for retrieval-augmented generation capabilities

Value Proposition

Amazon Neptune offers compelling advantages as a fully managed graph database service. It eliminates operational overhead by handling database management tasks like provisioning, patching, and backups automatically:

- Neptune provides superior performance with millisecond latency queries and the ability to handle billions of relationships
- The service offers flexibility with support for multiple graph models and query languages (Gremlin, openCypher, SPARQL), allowing customers to choose the best approach for their use cases
- Neptune's serverless option enables cost optimization by automatically scaling resources and charging only for actual usage
- The service integrates seamlessly with the AWS ecosystem and includes advanced features like machine learning capabilities, global database support for multi-region deployments, and built-in security features
- This combination of managed operations, high performance, flexibility, and AWS integration makes Neptune an ideal choice over self-managed graph databases or alternative solutions

Service Integration

Neptune integrates extensively with core AWS services to provide a comprehensive graph database solution. It seamlessly connects with Amazon S3 for data storage, backup, and bulk data loading operations:

- Integration with AWS Lambda enables serverless functions to interact with Neptune for real-time processing and event-driven architectures
- Amazon SageMaker AI integration powers Neptune ML capabilities for graph machine learning workflows

- Neptune works with Amazon VPC for network isolation and security, AWS IAM for access management, and AWS KMS for encryption key management
- The service integrates with Amazon CloudWatch for monitoring and metrics, AWS CloudTrail for auditing, and AWS Backup for automated backup management
- Data migration capabilities are provided through AWS Database Migration Service (DMS), while Amazon Kinesis enables real-time data streaming into Neptune
- Additional integrations include AWS Glue for data preparation, Amazon Athena for querying via connectors, and AWS API Gateway for building graph-based APIs

Onboarding and Offboarding

Getting started with Neptune is straightforward through multiple pathways. Customers can begin with the free trial offering 750 hours of db.t3.medium or db.t4g.medium instance usage, 10 million I/O requests, and 1 GB each of storage and backup storage for eligible new users:

- The AWS Management Console provides an intuitive interface to create Neptune clusters, while AWS CLI and APIs offer programmatic access
- Neptune Serverless allows immediate scaling without capacity planning, making it ideal for getting started quickly
- The service includes comprehensive documentation, getting started guides, and AWS Graph Notebook for learning and development
- Customers can use sample datasets and reference architectures from the AWS GitHub repository
- Training resources include AWS Skill Builder courses, workshops, and video tutorials
- The Neptune Workbench provides Jupyter notebooks for interactive development and visualization, enabling users to learn graph concepts and query languages without production environment costs

Getting Started Guide: https://docs.aws.amazon.com/neptune/latest/userguide/graph-get-started.html

Data Removal

Neptune provides comprehensive data removal options for offboarding. Customers can delete individual DB clusters which removes all data, automated backups, and cluster resources, though manual snapshots are preserved unless explicitly deleted. Point-in-time recovery data is retained for the backup retention period (1-35 days) before automatic deletion. For complete offboarding, customers should delete manual snapshots, remove any Neptune Analytics graphs, clean up associated AWS resources like VPC endpoints and Lambda functions, and ensure proper cleanup of integrated services. The service supports selective data deletion through standard graph query operations before cluster

termination. Cross-region replicas and global database configurations require separate deletion steps in each region.

Backup/Restore and Disaster Recovery

Neptune provides robust backup and disaster recovery capabilities through multiple mechanisms. Automated continuous incremental backups occur with configurable retention periods from 1 to 35 days, enabling point-in-time recovery to any second within the retention window. Manual database snapshots can be created and retained indefinitely for long-term backup needs. Neptune Global Database offers disaster recovery across AWS Regions with RPO of 1 second and RTO under 1 minute. Cross-region snapshot copying enables additional disaster recovery options. The service supports Neptune-to-Neptune replication using streams for real-time backup cluster synchronization. All backup operations occur without performance impact on production workloads, and storage-based replication ensures data durability across multiple Availability Zones.

Backup Capabilities

Neptune includes comprehensive built-in backup capabilities with automatic continuous incremental backups enabled by default. The service maintains backup retention periods from 1 to 35 days and supports manual snapshots for extended retention:

- Neptune integrates with AWS Backup for centralized backup management across multiple AWS services, allowing customers to create backup plans, automate backup schedules, and manage retention policies through a unified console
- Cross-region backup copying is supported for disaster recovery scenarios, and all backup operations are designed to have no performance impact on running database workloads

Disaster Recovery

Neptune supports disaster recovery through multiple approaches but does not directly integrate with AWS Elastic Disaster Recovery service. The primary disaster recovery mechanisms include Neptune Global Database for cross-region replication with sub-second RPO and sub-minute RTO, Neptune streams-based replication to maintain synchronized backup clusters in different regions, and cross-region snapshot copying for recovery scenarios:

- Managed planned failover capabilities allow controlled region switching for maintenance operations, while unplanned failover supports emergency recovery situations
- These native Neptune disaster recovery features provide comprehensive protection without requiring AWS Elastic Disaster Recovery integration

Recovery Objectives

Neptune helps customers achieve aggressive RPO and RTO objectives through its disaster recovery features. Neptune Global Database delivers RPO of 1 second and RTO of less than 1 minute for cross-region scenarios using storage-based replication. For local high availability, Multi-AZ deployments with read replicas provide automatic failover typically within minutes. Point-in-time recovery enables restoration to any second within the backup retention period. Neptune streams replication can achieve RPO and RTO measured in minutes for custom disaster recovery implementations. The combination of continuous backups, automated failover, and global database capabilities allows customers to design solutions meeting strict availability and recovery requirements for mission-critical graph database workloads.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/neptune/latest/userguide/limits.html

FAQ: https://aws.amazon.com/neptune/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/neptune/

User Guide: https://docs.aws.amazon.com/neptune/latest/userguide/

API Reference: https://docs.aws.amazon.com/neptune/latest/apiref/

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/neptune/pricing/

Cost Optimization

Neptune offers several unique cost optimization options tailored for graph database workloads. Neptune Serverless provides automatic scaling with pay-per-use pricing, potentially reducing costs by up to 90% for variable workloads by charging only for active capacity:

- Customers can choose between Neptune Standard for typical I/O usage or Neptune I/O-Optimized for predictable pricing on I/O-intensive applications
- The service supports both on-demand and reserved instance pricing through Database Savings Plans with 1-year commitments
- AWS Graviton4 R8g instances offer up to 4.7 times better price-performance for write workloads compared to previous generations

- Neptune Analytics allows pausing and resuming graph workloads with reduced compute pricing when paused
- Query optimization through proper indexing and graph model design can significantly reduce I/O costs, while strategic use of read replicas can distribute load and optimize resource utilization

Savings Considerations

Key cost savings strategies for Neptune include leveraging Neptune Serverless for variable workloads to avoid over-provisioning and paying only for actual usage. Database Savings Plans provide significant discounts for predictable workloads with 1-year commitments:

- Upgrading to newer instance generations like Graviton4 R8g instances can deliver substantial price-performance improvements - up to 4.7 times better for write operations and 3.7 times better for read operations
- Right-sizing instances based on actual performance metrics like MainRequestQueuePendingRequests and BufferCacheHitRatio prevents over-provisioning
- Optimizing graph data models and queries reduces I/O transfer costs, while storing large objects outside Neptune in S3 can minimize storage expenses
- Strategic use of read replicas for load distribution and choosing appropriate backup retention periods help manage overall costs
- For development environments, utilizing the free trial and properly scheduling instance start/stop operations can provide additional savings

Service Name

Amazon OpenSearch Service

Service Overview

Amazon OpenSearch Service is a fully managed service that simplifies the deployment, operation, and scaling of OpenSearch clusters in the AWS Cloud. OpenSearch is a fully open-source search and analytics engine used for log analytics, real-time application monitoring, and clickstream analysis. Amazon OpenSearch Service provisions all resources for an OpenSearch cluster and manages its infrastructure, including automatic node replacement and scaling. The service offers built-in integrations with OpenSearch Dashboards, Logstash, and AWS services including Amazon Data Firehose, AWS Lambda, and Amazon CloudWatch, making it ideal for querying and searching large amounts of data such as analyzing activity logs, CloudWatch Logs, product usage data, social media sentiments, and data stream updates from other AWS services.

Key Use Cases

- Log analytics: Processing and analyzing machine-generated time series data to gain insights into operations, security, and user behavior from application logs, system telemetry, and operational data
- Real-time application monitoring: Monitoring application performance, identifying and troubleshooting errors, tracking system health, and analyzing user behavior patterns in real-time
- Full-text search: Building sophisticated search applications with capabilities like prefix match, multi-match, wildcard match, and range filtering for e-commerce, content management, and document search
- Security analytics: Analyzing security logs, detecting anomalies, monitoring threats, and performing security investigations with built-in security rules and custom detection capabilities
- Business intelligence and analytics: Creating interactive dashboards, performing data visualization, analyzing clickstream data, and deriving business insights from large datasets
- Observability and monitoring: Handling time-series data from logs, metrics, and traces for DevOps teams to identify performance bottlenecks and system anomalies

Features

- **Serverless Deployment**: On-demand, auto-scaling configuration that automatically scales compute resources based on application needs with cost-effective pricing model

- **Multi-tier Storage**: UltraWarm for cost-effective storage of infrequently accessed data using S3 and caching, and cold storage for rarely accessed data with significant cost savings
- **Fine-grained Access Control**: Role-based access control, security at the index, document, and field level, multi-tenancy support, and HTTP basic authentication
- **Vector Search Engine**: Rich lexical and vector search capabilities for AI applications with GPU-accelerated and auto-optimized vector indexes supporting billion-scale vector databases
- **Multi-AZ with Standby**: Enhanced availability and performance configuration ensuring resilience to single AZ failures with 99.99% availability SLA
- **Index State Management**: Automated lifecycle management for indexes including transitions between hot, warm, and cold storage tiers based on customizable policies
- **OpenSearch Dashboards**: Native visualization tool for creating interactive dashboards, exploring data, performing analytics, and building custom visualizations without additional fees
- **Cross-cluster Operations**: Cross-cluster search and replication capabilities for disaster recovery, reduced latency, and querying data across multiple domains
- **Automated Snapshots**: Hourly automated snapshots stored in S3 for data durability and manual snapshot capabilities for cross-region disaster recovery
- **Agentic Search**: Intelligent agent-driven system that understands user intent, orchestrates tools, generates OpenSearch DSL queries, and provides natural language interactions

Value Proposition

Amazon OpenSearch Service provides a fully managed infrastructure that eliminates the complexity and operational overhead of managing OpenSearch clusters while delivering enterprise-grade capabilities. Unlike self-managed solutions, it offers automatic hardware provisioning, software installation and patching, 24x7 monitoring and repairs, zero downtime updates, and seamless scaling up to 3 petabytes of data:

- The service provides significant cost advantages through tiered storage options like UltraWarm (up to 40% cost savings) and cold storage, AWS Graviton2 instances offering up to 44% price/performance improvements, and Reserved Instances with up to 52% discounts
- Enterprise-grade security is built-in with encryption at rest and in transit, VPC integration, fine-grained access control, and compliance with standards including HIPAA, SOC 2, PCI DSS, and FedRAMP
- The service offers unique capabilities like serverless deployment with automatic scaling, vector search for AI applications, Multi-AZ with Standby for 99.99% availability, and native integrations with the broader AWS ecosystem including CloudWatch, Lambda, S3, and Kinesis

Service Integration

Amazon OpenSearch Service integrates deeply with core AWS services to provide a comprehensive analytics and search platform. Data ingestion is supported through Amazon Kinesis Data Streams, Amazon Data Firehose, AWS Lambda, and direct integration with Amazon S3 for log analytics:

- Monitoring and observability are enhanced through native Amazon CloudWatch integration for metrics and alerting, plus AWS CloudTrail for API call auditing
- Security integrations include AWS IAM for access control, AWS Key Management Service for encryption key management, and Amazon VPC for network isolation
- For data processing, the service connects with Amazon DynamoDB Streams for real-time data changes, Amazon SNS for notifications, and AWS Glue for ETL operations
- Storage integration with Amazon S3 enables UltraWarm and cold storage tiers, while Amazon Security Lake integration provides advanced security analytics
- The service also supports Amazon Cognito for authentication, AWS Organizations for multi-account management, and AWS Service Catalog for standardized deployments, creating a unified experience across the AWS ecosystem

Onboarding and Offboarding

Getting started with Amazon OpenSearch Service involves creating an AWS account, setting up appropriate IAM permissions, and choosing between managed clusters or serverless collections based on workload requirements. Customers can create their first domain through the OpenSearch Service console, AWS CLI, or AWS SDKs, specifying domain name, OpenSearch version, instance types, storage options, and network configuration:

- The service provides Migration Assistant to facilitate migrations from self-managed Elasticsearch or OpenSearch clusters, automating data transfer and optimizing performance
- Initial setup includes configuring security settings such as fine-grained access control, encryption options, and VPC settings
- Data can be ingested through various methods including direct API calls, Amazon Kinesis, AWS Lambda, or bulk upload operations
- OpenSearch Dashboards is automatically available for visualization and data exploration
- AWS provides comprehensive tutorials, workshops, and documentation through AWS Skill Builder, plus migration support and training for eligible customers to ensure successful adoption

Getting Started Guide: https://docs.aws.amazon.com/opensearch-service/latest/developerguide/gsg.html

Data Removal

When offboarding from Amazon OpenSearch Service, customers must first export or migrate their data using manual snapshots stored in their own S3 buckets for cross-region or cross-account transfers. All domain data, including indexes and metadata, should be backed up before deletion. The offboarding process involves deleting the OpenSearch Service domain through the console, CLI, or API, which permanently removes all associated resources including compute instances, storage volumes, and automated snapshots. Manual snapshots stored in customer-managed S3 buckets remain accessible after domain deletion, enabling data recovery if needed. Customers should also remove associated IAM policies, security groups, and VPC configurations as part of the cleanup process.

Backup/Restore and Disaster Recovery

Amazon OpenSearch Service provides comprehensive backup and disaster recovery through automated and manual snapshot capabilities. The service automatically creates hourly snapshots of all domains, storing them in Amazon S3 for up to 14 days at no additional charge. Manual snapshots can be created and stored in customer-managed S3 buckets for longer retention and cross-region disaster recovery. Snapshots include index metadata, cluster metadata, and index data, enabling complete domain restoration. For disaster recovery, customers can implement active-passive strategies using cross-region snapshot restoration, with Amazon Route 53 handling traffic redirection during failover scenarios. The service also supports cross-cluster replication for real-time data synchronization across geographically distributed domains.

Backup Capabilities

Amazon OpenSearch Service includes built-in automated backup through hourly snapshots stored in Amazon S3, retained for 14 days at no additional cost. Manual snapshots can be created on-demand and stored in customer-managed S3 buckets for extended retention and cross-region access:

- The service supports Index State Management policies for automated snapshot creation of critical indices
- However, Amazon OpenSearch Service does not directly integrate with AWS Backup service, requiring customers to use native OpenSearch snapshot mechanisms for backup and restore operations

Disaster Recovery

Amazon OpenSearch Service supports disaster recovery through multiple mechanisms including cross-region snapshot restoration, cross-cluster replication, and Multi-AZ deployments with standby instances. The service enables active-passive disaster recovery strategies using manual snapshots stored in S3 buckets across regions:

- Cross-cluster replication provides real-time data synchronization between domains in different regions
- However, Amazon OpenSearch Service does not integrate with AWS Elastic Disaster Recovery service, instead relying on OpenSearch-native capabilities and AWS infrastructure resilience for disaster recovery scenarios

Recovery Objectives

Amazon OpenSearch Service helps customers achieve their RPO and RTO objectives through automated hourly snapshots (RPO of 1 hour), Multi-AZ with Standby configuration for near-zero RTO during single AZ failures, and cross-cluster replication for real-time data synchronization. Manual snapshots enable customer-defined RPO based on backup frequency. The service's blue-green deployment model ensures zero downtime for most configuration changes. For critical workloads, the Multi-AZ with Standby feature provides 99.99% availability SLA and rapid failover capabilities, while snapshot-based recovery can restore entire domains within minutes to hours depending on data volume.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/general/latest/gr/opensearch-service.html

FAQ: https://aws.amazon.com/opensearch-service/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/opensearch-service/latest/developerguide/what-is.html

User Guide: https://docs.aws.amazon.com/opensearch-service/latest/developerguide/

API Reference: https://docs.aws.amazon.com/opensearch-service/latest/APIReference/

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/opensearch-service/pricing/

Cost Optimization

Amazon OpenSearch Service offers unique cost optimization through tiered storage architecture with UltraWarm providing up to 40% cost savings for infrequently accessed data using S3-based storage, and cold storage for rarely accessed data at the lowest cost. AWS Graviton2 instances deliver up to 44% better price/performance compared to previous generation instances:

- Reserved Instances provide up to 52% discounts for predictable workloads with one or three-year terms
- The serverless deployment option enables pay-per-use pricing, eliminating overprovisioning costs for variable workloads
- OpenSearch Optimized Instances (OR1 and OI2) offer improved durability and performance with cost savings
- Customers can implement Index State Management policies to automatically transition data through storage tiers based on age and access patterns, optimizing costs while maintaining performance for active data

Savings Considerations

Primary cost savings opportunities include implementing storage tier strategies by moving infrequently accessed data to UltraWarm (40% savings) and cold storage for maximum cost reduction. Reserved Instances provide significant discounts of up to 52% for predictable workloads compared to on-demand pricing:

- Choosing appropriate instance types including AWS Graviton2 instances for better price/performance ratios and right-sizing clusters based on actual usage patterns prevents overprovisioning
- Serverless deployment eliminates idle resource costs for variable workloads
- Automated Index State Management reduces manual operational overhead while optimizing storage costs
- Regular monitoring using CloudWatch metrics and AWS Cost Explorer helps identify optimization opportunities
- Leveraging free automated snapshots instead of manual snapshots reduces storage costs, while cross-region data transfer optimization minimizes bandwidth charges for multi-region deployments

Service Name

Amazon Personalize

Service Overview

Amazon Personalize is a fully managed machine learning service that generates item recommendations and user segments based on user interaction data. It enables developers to build personalized experiences for applications including video streaming, e-commerce, email marketing, and search results. The service uses historical data and real-time events to train machine learning models that predict user preferences and behaviors. Amazon Personalize offers both preconfigured domain-specific recommenders for VIDEO_ON_DEMAND and ECOMMERCE use cases, as well as customizable solutions for unique business requirements. It provides real-time recommendation APIs and batch inference capabilities for scalable personalized experiences.

Key Use Cases

- Personalizing video streaming apps with recommendations like Top picks for you, More like X, and Most popular videos
- Adding product recommendations to e-commerce applications including Recommended for you and Frequently bought together
- Creating targeted marketing campaigns and personalized emails using user segments based on item affinity
- Implementing real-time next best action recommendations for user engagement like loyalty program enrollment
- Personalizing search results by re-ranking items based on user preferences and behavior patterns

Features

- **Real-time Personalization**: Updates recommendations based on evolving user interests by recording interactions with items or actions
- **Domain Recommenders**: Preconfigured resources optimized for VIDEO_ON_DEMAND and ECOMMERCE domains with use-case specific recipes
- **Custom Solutions**: Configurable machine learning solutions using customizable recipes for unique business requirements
- **Batch Recommendations**: Bulk recommendation generation for email campaigns and offline processing workflows
- **User Segmentation**: Generate segments of users based on affinity for items or metadata for targeted marketing
- **Automatic Training**: Automatically retrain solutions every 7 days using latest data to maintain recommendation relevance

- **Exploration**: Introduces less likely items to improve discovery and engagement with configurable exploration weights
- **Generative AI Integration**: Content Generator adds descriptive themes to batch recommendations using generative AI
- **Rule-based Promotions**: Support for promoting specific items in recommendations based on business rules

Value Proposition

Amazon Personalize provides fully managed machine learning capabilities without requiring ML expertise, offering faster time-to-market for personalized experiences. It automatically handles model training, hyperparameter optimization, and infrastructure scaling, reducing operational overhead:

- The service supports catalogs with up to 5 million items through new v2 recipes with lower inference latency
- Built-in integration with AWS services like S3, CloudWatch, and Amplify simplifies data workflows
- Real-time personalization updates recommendations immediately based on user behavior
- The flexible pricing model with no minimum fees and AWS Free Tier support makes it accessible for businesses of all sizes
- Advanced features like generative AI themes, rule-based promotions, and automatic solution training provide competitive advantages in delivering sophisticated personalized experiences

Service Integration

Amazon Personalize integrates deeply with core AWS services for comprehensive personalized application development. It connects with Amazon S3 for bulk data import and export, Amazon SageMaker AI Data Wrangler for importing data from 40+ sources, and AWS Amplify for frontend integration:

- CloudWatch integration provides monitoring and metrics, while CloudWatch Evidently enables A/B testing of recommendations
- The service works with AWS Lambda for real-time event processing and Step Functions for workflow orchestration
- Integration with Amazon Bedrock enables generative AI content creation using recommendation metadata
- IAM provides fine-grained access control, and AWS PrivateLink supports VPC endpoints for secure connectivity
- OpenSearch integration allows for personalized search result ranking

Onboarding and Offboarding

Customers begin by choosing between Domain dataset groups for VIDEO_ON_DEMAND or ECOMMERCE applications, or Custom dataset groups for unique use cases. The process involves preparing training data in CSV format, creating schemas to define data structure, and establishing a dataset group container:

- Data import requires uploading files to S3 and creating dataset import jobs
- For Domain groups, customers create recommenders; for Custom groups, they build solutions and deploy campaigns
- AWS provides comprehensive getting started tutorials using console, CLI, or SDKs in Python, Java, and JavaScript
- The Magic Movie Machine interactive learning experience helps first-time users understand features
- Multiple SDK examples and Jupyter notebooks are available in the amazon-personalize-samples GitHub repository

Getting Started Guide: https://docs.aws.amazon.com/personalize/latest/dg/getting-started.html

Data Removal

Offboarding requires systematic deletion of resources in specific order: campaigns or recommenders first, then solutions, event trackers, metric attributions, filters, datasets, and finally dataset groups and schemas. Data deletion jobs can remove specific users and their interaction data using CSV files with user IDs stored in S3. The process involves creating data deletion jobs that can take up to 24 hours to complete. All associated resources must be in ACTIVE status with no pending operations before deletion. Resource deletion is irreversible, requiring careful planning to avoid data loss.

Backup/Restore and Disaster Recovery

Amazon Personalize does not provide native backup capabilities or integration with AWS Backup service. Data durability relies on AWS infrastructure resilience and multi-AZ deployment of the underlying managed service. Customers are responsible for backing up their source data in S3 and maintaining copies of schemas, dataset configurations, and solution parameters. Recovery involves recreating dataset groups, reimporting data, and retraining models. Historical interaction data and model configurations should be preserved externally for disaster recovery scenarios.

Backup Capabilities

Amazon Personalize does not offer built-in backup capabilities or integration with AWS Backup service. The service relies on AWS managed infrastructure for data durability but does not provide customer-accessible backup features. Customers must implement their

own data protection strategies by maintaining copies of training data, schemas, and configuration parameters in external storage like S3 for recovery purposes.

Disaster Recovery

Amazon Personalize does not integrate with AWS Elastic Disaster Recovery service. The fully managed service handles infrastructure resilience automatically, but customers cannot directly control disaster recovery processes:

- Recovery from regional failures requires recreating resources in alternate regions, reimporting data, and retraining models
- Cross-region data replication must be implemented by customers using S3 cross-region replication for training datasets

Recovery Objectives

Amazon Personalize supports business continuity through automatic infrastructure management and multi-AZ deployment, but specific RTO/RPO targets depend on customer implementation. Recovery time is primarily determined by model retraining duration, which varies based on data volume and complexity. RPO depends on real-time event streaming frequency and data backup strategies implemented by customers. The service's managed nature provides inherent resilience, but customers must plan for data recreation and model retraining in disaster scenarios.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/personalize/latest/dg/limits.html

FAQ: https://aws.amazon.com/personalize/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/personalize/latest/dg/

User Guide: https://docs.aws.amazon.com/personalize/latest/dg/what-is-personalize.html

API Reference: https://docs.aws.amazon.com/personalize/latest/dg/API_Reference.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/personalize/pricing/

Cost Optimization

Amazon Personalize offers several cost optimization strategies including tiered pricing for recommendations with volume discounts at 72 million and 720 million request thresholds. Customers can optimize costs by choosing appropriate training frequencies, using batch recommendations for bulk operations instead of real-time requests, and leveraging the AWS Free Tier for initial 2 months:

- Automatic training can be disabled to control training costs while maintaining model performance
- Data processing costs can be minimized by optimizing data formats and import frequency
- Use-case optimized recommenders provide cost-effective solutions for common scenarios without custom development overhead

Savings Considerations

Primary cost savings come from the fully managed service model eliminating infrastructure and ML expertise requirements, reducing total cost of ownership. The pay-as-you-use pricing with no upfront commitments allows cost scaling with business growth:

- Volume-based tiered pricing provides automatic discounts at higher usage levels
- Batch processing for bulk recommendations costs significantly less than individual real-time requests
- AWS Free Tier provides substantial cost savings during initial evaluation and small-scale implementations
- Proper data preparation and schema design can reduce unnecessary data processing costs
- Using domain-specific recommenders instead of custom solutions can lower development and maintenance costs while achieving faster implementation

Service Name

Amazon Polly

Service Overview

Amazon Polly is a cloud-based text-to-speech (TTS) service offered by AWS that converts text into lifelike speech using advanced machine learning technology. The service enables developers to create engaging and accessible applications across multiple platforms including mobile apps, eLearning platforms, accessibility applications, and IoT devices. Amazon Polly supports over 60 voices in multiple languages and offers four voice engines: Standard, Neural, Long-form, and Generative, each optimized for different use cases and speech quality requirements.

Key Use Cases

- Mobile applications: Adding voice capabilities to mobile apps for enhanced user experiences and accessibility
- eLearning platforms: Converting educational content to speech for auditory learners and accessibility compliance
- Accessibility applications: Creating inclusive experiences for visually impaired users through document-to-speech conversion
- IoT devices: Enabling voice responses and notifications in smart home and embedded systems
- Contact centers: Integrating with Amazon Connect for interactive voice response (IVR) systems
- Content creation: Generating audio narration for videos, podcasts, and marketing materials

Features

- **Multiple Voice Engines**: Standard, Neural, Long-form, and Generative voices with varying quality levels and pricing
- **SSML Support**: Speech Synthesis Markup Language for controlling speech rate, pitch, emphasis, and pronunciation
- **Speech Marks**: Metadata providing timing information for synchronizing speech with visual elements
- **Custom Lexicons**: Ability to modify pronunciation of specific words and acronyms
- **Brand Voice**: Custom Neural Text-to-Speech voices for organizations
- **Multiple Audio Formats**: Support for MP3, OGG Vorbis, and raw PCM output formats

- **Newscaster Speaking Style**: Professional news-style delivery for US English, British English, and US Spanish voices
- **Time-driven Prosody**: Ability to define maximum speech duration for localization and dubbing

Value Proposition

Amazon Polly provides superior value through its pay-as-you-go pricing model eliminating upfront costs and setup fees, allowing organizations to start small and scale as needed. The service delivers high-quality, natural-sounding speech using advanced neural technology with fast response times suitable for real-time applications:

- Cloud-based deployment reduces local resource requirements while providing instant access to all voices and languages with automatic improvements
- Amazon Polly offers extensive customization options including custom lexicons and brand voices, supports caching and replay at no additional cost, and maintains compliance certifications for HIPAA and PCI DSS regulated workloads

Service Integration

Amazon Polly integrates natively with Amazon Connect for contact center solutions and Amazon Chime SDK for communication applications. The service works seamlessly with AWS Lambda for serverless processing workflows, Amazon S3 for audio storage and retrieval, and AWS Step Functions for orchestrating complex text-to-speech pipelines:

- Additional integrations include Amazon Textract for document-to-speech conversion, Amazon Translate for multilingual workflows, Amazon API Gateway for REST API access, and Amazon CloudWatch for monitoring and analytics
- Amazon Polly also supports integration with third-party contact center platforms like Genesys Cloud CX

Onboarding and Offboarding

Getting started with Amazon Polly requires signing up for an AWS account, which automatically provides access to the service with charges only for actual usage. Customers can begin immediately using the AWS Management Console, AWS CLI, or SDKs in multiple programming languages including Java, Python, Node.js, .NET, PHP, Ruby, Go, and C++:

- The service offers comprehensive getting started documentation, sample code, and tutorials for various platforms including iOS and Android mobile development
- New AWS customers receive free tier benefits with varying character limits for each voice type during the first 12 months

Getting Started Guide: https://docs.aws.amazon.com/polly/latest/dg/getting-started.html

Data Removal

Amazon Polly does not retain text inputs processed by the service, ensuring automatic data removal upon processing completion. Users maintain full ownership of their content and generated audio files. For offboarding, customers simply stop using the service as there are no long-term commitments or stored data requiring manual deletion. Generated speech can be cached and replayed indefinitely without additional charges, but storage and management of audio files remain the customer's responsibility.

Backup/Restore and Disaster Recovery

Amazon Polly operates as a fully managed service with built-in resilience across multiple AWS Availability Zones. The service does not store customer text inputs, eliminating the need for traditional backup of processed content. Customers are responsible for backing up their audio files stored in services like Amazon S3. The service maintains high availability through AWS's global infrastructure with automatic failover capabilities built into the platform architecture.

Backup Capabilities

Amazon Polly does not provide built-in backup capabilities as it does not retain customer text inputs or generated audio files. The service is stateless, processing text requests and returning audio streams without persistent storage:

- Customers must implement their own backup strategies for generated audio content using services like Amazon S3 with versioning and cross-region replication
- Amazon Polly does not integrate with AWS Backup service due to its stateless nature

Disaster Recovery

Amazon Polly supports disaster recovery through its multi-region availability and stateless architecture. The service operates across multiple AWS regions, allowing customers to redirect traffic to alternative regions during outages:

- Amazon Polly does not integrate with AWS Elastic Disaster Recovery as it does not maintain persistent infrastructure requiring replication
- The service's cloud-native design provides inherent resilience with automatic geographic distribution of processing capabilities

Recovery Objectives

Amazon Polly helps customers achieve low Recovery Point Objectives (RPO) and Recovery Time Objectives (RTO) through its stateless design and multi-region availability. Since no customer data is retained, RPO is essentially zero. RTO is minimized through the service's ability to process requests in multiple regions simultaneously. Customers can implement

automated failover using Amazon Route 53 health checks to redirect traffic to healthy regions, typically achieving RTOs of minutes rather than hours.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/polly/latest/dg/limits.html

FAQ: https://aws.amazon.com/polly/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/polly/latest/dg/what-is.html

User Guide: https://docs.aws.amazon.com/polly/latest/dg/getting-started.html

API Reference: https://docs.aws.amazon.com/polly/latest/dg/API_Reference.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/polly/pricing/

Cost Optimization

Amazon Polly offers several cost optimization strategies through its pay-per-character pricing model. Customers can optimize costs by selecting appropriate voice engines based on quality requirements - Standard voices at $4.00 per million characters for basic needs, versus Neural voices at $16.00 for higher quality:

- Caching generated audio eliminates repeated synthesis costs
- Using asynchronous processing for long-form content can be more cost-effective than real-time synthesis
- Implementing efficient text preprocessing to remove unnecessary characters and using SSML judiciously helps minimize billable character counts while maintaining speech quality

Savings Considerations

Primary cost savings come from the pay-as-you-go model eliminating upfront licensing fees and infrastructure costs. Caching synthesized speech allows unlimited replay without additional charges, significantly reducing costs for frequently accessed content:

- Choosing the most cost-effective voice engine for specific use cases - Standard voices for basic applications versus premium Neural voices only when necessary - can reduce costs by up to 75%

- Efficient application design with proper error handling and retry logic minimizes unnecessary API calls
- Batch processing longer content using asynchronous APIs can be more economical than multiple real-time requests

Service Name

Amazon Quick Suite

Service Overview

Amazon Quick Suite is an agentic AI-powered workspace that combines research, business intelligence, and automation capabilities into a unified digital workspace. It helps users retrieve insights from business data across the internet and their documents, including third-party applications, databases, and other data sources. Quick Suite seamlessly transitions from getting answers to taking action in applications like Jira and ServiceNow, and can automate tasks from routine duties to complex business processes. The service keeps data private and allows users to tailor their experience while providing AI-powered teammates for comprehensive business productivity.

Key Use Cases

- Research and Analysis: Conducting comprehensive research across enterprise data and external sources to deliver contextualized, actionable insights in minutes
- Business Intelligence: Transforming data into actionable insights through natural language queries and interactive visualizations with Quick Sight
- Process Automation: Automating repetitive tasks and complex business processes from simple workflows to multi-department enterprise-scale automation
- Data Consolidation: Creating unified knowledge foundations that consolidate documents, files, and application data for AI-driven insights
- Cross-Application Integration: Taking actions in business applications like creating Jira tickets or ServiceNow incidents directly from insights

Features

- **Quick Index**: Unified knowledge foundation that consolidates documents, files, and application data to power AI-driven insights and responses
- **Quick Research**: Powerful agent that conducts comprehensive research across enterprise data and external sources to deliver contextualized insights
- **Quick Sight**: AI-powered business intelligence solution that transforms data into actionable insights through natural language queries and interactive visualizations
- **Quick Flows**: Tool that allows users to automate repetitive tasks by describing workflows using natural language
- **Quick Automate**: Enterprise-scale process automation that transforms complex business processes into multi-agent workflows
- **Spaces**: Provides a straightforward way for users to add context and organize work within the platform

Value Proposition

Amazon Quick Suite provides a unified agentic AI workspace that eliminates the need to switch between multiple applications for data gathering, analysis, and task execution. Unlike fragmented traditional approaches, Quick Suite combines research, business intelligence, and automation in one platform, reducing manual work and enabling faster decision-making. The service offers secure, private data handling where queries and data are never used to train models, while providing AI-powered teammates that can handle everything from simple tasks to complex multi-department workflows, ultimately improving productivity and reducing operational overhead.

Service Integration

Amazon Quick Suite integrates deeply with AWS services including Amazon Q Business for generative AI capabilities, Amazon QuickSight for business intelligence functionality, AWS Lambda and SNS for automation workflows, and IAM Identity Center for authentication. The service connects to various AWS data sources including Amazon S3, Amazon Redshift, Amazon RDS, and other databases. Quick Suite also integrates with third-party applications like Jira and ServiceNow for action execution, and utilizes AWS's global infrastructure for high availability and security compliance including HIPAA, SOC 2, and PCI DSS certifications.

Onboarding and Offboarding

Getting started with Amazon Quick Suite requires AWS administrator activation in just a few steps through the AWS Management Console. New customers receive a 30-day free trial for up to 25 users:

- The service supports IAM Identity Center for authentication and can be quickly deployed without infrastructure setup
- Users can immediately begin uploading documents to Quick Index, creating analyses through natural language queries, and setting up automation workflows
- The platform provides guided tutorials and sample data to help users explore capabilities across research, business intelligence, and automation features

Getting Started Guide: https://docs.aws.amazon.com/quicksuite/latest/userguide/quick-sight-getting-started.html

Data Removal

Amazon Quick Suite maintains data privacy with user queries and data never used to train models. For offboarding, customers can remove their data through standard AWS data deletion practices. The service provides secure data handling with encryption at rest and in transit, and follows AWS's standard data retention and deletion policies for managed services.

Backup/Restore and Disaster Recovery

Amazon Quick Suite leverages AWS's cloud infrastructure for high availability and durability. As a fully managed service, AWS handles backup and disaster recovery operations automatically. Data is replicated across multiple Availability Zones within regions for resilience, and the service maintains business continuity through AWS's proven disaster recovery capabilities.

Backup Capabilities

Amazon Quick Suite provides automatic backup capabilities as part of its managed service infrastructure. The service maintains data persistence through AWS's underlying storage systems with automatic replication and point-in-time recovery capabilities built into the platform architecture.

Disaster Recovery

Amazon Quick Suite supports disaster recovery through AWS's multi-region infrastructure and automatic failover capabilities. The service is designed with high availability and can recover from regional outages through AWS's proven disaster recovery frameworks and cross-region replication strategies.

Recovery Objectives

Amazon Quick Suite helps customers meet RPO and RTO objectives through its serverless, highly available architecture built on AWS's resilient infrastructure. The service automatically scales and provides continuous availability, supporting business continuity requirements with minimal downtime during recovery scenarios.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/general/latest/gr/quicksuite.html

FAQ: https://aws.amazon.com/quicksuite/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/quicksuite/latest/userguide/

User Guide: https://docs.aws.amazon.com/quicksuite/latest/userguide/

API Reference: https://docs.aws.amazon.com/quicksuite/latest/api/

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/quicksuite/pricing/

Cost Optimization

Amazon Quick Suite offers per-user subscription-based pricing with consumption-based charges for optional features like Quick Index, allowing organizations to scale costs with actual usage. The service eliminates infrastructure management overhead and reduces operational costs by consolidating research, business intelligence, and automation tools into a single platform. Organizations can optimize costs by leveraging the 30-day free trial, carefully managing user subscriptions, and monitoring consumption-based feature usage to align with business needs.

Savings Considerations

Amazon Quick Suite delivers cost savings by replacing multiple point solutions for research, business intelligence, and automation with a single integrated platform, reducing software licensing costs and operational overhead. The serverless architecture eliminates infrastructure management costs, while the AI-powered automation capabilities reduce manual labor costs across business processes. Organizations can achieve significant productivity gains by reducing time spent switching between applications and manually gathering data, with the potential for substantial ROI through improved decision-making speed and process automation efficiency.

Service Name

Amazon Q Business

Service Overview

Amazon Q Business is a fully managed, generative AI-powered assistant designed for enterprises that uses advanced natural language processing to answer questions, provide summaries, generate content, and automate tasks based on enterprise data. It employs Retrieval Augmented Generation (RAG) technology to deliver permissions-aware responses from multiple data sources with citations. The service handles machine learning infrastructure and models automatically, providing accurate answers by analyzing information across all enterprise content while respecting existing access controls and security permissions.

Key Use Cases

- Knowledge management and support: Enterprise-wide information retrieval and institutional knowledge sharing across business domains
- Employee onboarding and training: Streamlined access to training materials, policies, and procedural documentation for new hires
- IT help desk support: Automated troubleshooting and technical support through conversational AI interface
- Content creation and document automation: Generate summaries, outlines, drafts, and marketing materials from enterprise data
- Workflow automation and task completion: Create custom Amazon Q Apps to automate repetitive business processes and integrate with third-party applications

Features

- **Agentic Retrieval Augmented Generation (RAG)**: Advanced NLP technique using multiple intelligent agents and tools to process queries and retrieve responses with continuous quality monitoring
- **Source Attribution with Citations**: Provides source attributions and citations to verify responses and ensure information transparency
- **40+ Pre-built Data Connectors**: Built-in connectors for enterprise applications including Salesforce, SharePoint, Confluence, Jira, and Amazon S3
- **Amazon Q Apps**: Lightweight, purpose-built applications within the web experience for specific tasks and workflow automation
- **Chat Orchestration and Plugins**: Automatically manages chat requests across configured plugins and data sources, supporting 50+ business application integrations

- **Guardrails and Topic Controls**: Global and topic-level controls including PII detection, blocked topics, and content filtering for security and compliance
- **Document Enrichment and OCR**: Custom document content manipulation, OCR capabilities for scanned PDFs, and multimedia content extraction
- **Response Personalization**: Customizes chat responses based on end user metadata and job-related information from identity providers

Value Proposition

Amazon Q Business offers a unified, fully managed generative AI solution that eliminates the complexity of building and maintaining AI infrastructure while providing enterprise-grade security and compliance. It reduces time-to-value by offering pre-built connectors to over 40 data sources, built-in indexing pipelines, and ready-to-deploy interfaces:

- The service respects existing access controls and permissions, ensuring secure data handling while enabling employees to quickly find information, automate workflows, and increase productivity
- Its subscription-based pricing model with transparent costs makes it cost-effective compared to building custom AI solutions, while features like source citations and guardrails ensure reliable, trustworthy responses for business-critical applications

Service Integration

Amazon Q Business integrates deeply with core AWS services including AWS IAM Identity Center for user authentication and authorization, Amazon Bedrock for foundation model access, Amazon S3 for data storage and retrieval, Amazon Kendra as an alternative retriever option, and AWS Lambda for custom document processing logic. It leverages Amazon AppFlow to pull data from unsupported sources, integrates with AWS Step Functions for data orchestration, and uses Amazon CloudWatch for logging and monitoring. The service also supports AWS CloudTrail for audit logging and works within the broader Quick Suite ecosystem including Amazon QuickSight for enhanced analytics and insights across enterprise data sources.

Onboarding and Offboarding

Customers begin by setting up an AWS account and configuring AWS IAM Identity Center or IAM for user management, followed by creating required IAM roles for service permissions. The setup process involves creating an Amazon Q Business application environment through the AWS Management Console, selecting subscription tiers (Lite or Pro), and configuring index types (Starter or Enterprise):

- Next, administrators add data sources using pre-built connectors, configure user access and groups, and deploy the web experience interface
- The service provides guided setup wizards, sample applications, and starter kits to accelerate implementation, with optional steps including custom guardrails

configuration, plugin enablement, and integration with third-party applications like Slack or Microsoft Teams

Getting Started Guide: https://docs.aws.amazon.com/amazonq/latest/qbusiness-ug/getting-started.html

Data Removal

Amazon Q Business provides data removal capabilities through administrative controls and API operations. Organizations can delete conversation history, remove indexed documents through data source disconnection and synchronization, and delete entire application environments through the console or API. All user data including prompts, responses, and uploaded documents can be permanently removed. The service automatically saves conversation history for one month with administrator ability to delete earlier. Complete offboarding involves removing user subscriptions, disconnecting data sources, and deleting the application environment and associated indexes.

Backup/Restore and Disaster Recovery

Amazon Q Business operates as a fully managed service where AWS handles infrastructure backup and disaster recovery automatically across multiple Availability Zones. The service maintains data durability through AWS's underlying storage systems, but customers are responsible for backing up their source data in connected systems. Index data can be reconstructed through data source resynchronization, and application configurations can be recreated through Infrastructure as Code templates. The service provides high availability within supported regions but does not offer built-in cross-region replication capabilities.

Backup Capabilities

Amazon Q Business does not provide traditional backup capabilities as it operates as a fully managed service with automatic data durability. The service can reconstruct indexes by resyncing from original data sources, and conversation history is temporarily stored with configurable retention:

- AWS manages underlying infrastructure backup automatically, but customers should maintain backups of their source data systems
- Application configurations and settings can be preserved through CloudFormation templates or API-based configuration management for environment recreation purposes

Disaster Recovery

Amazon Q Business supports disaster recovery through its fully managed, multi-AZ architecture within supported regions, providing automatic failover capabilities for infrastructure components. The service does not integrate directly with AWS Elastic

Disaster Recovery, but offers resilience through data source redundancy and the ability to quickly recreate application environments:

- Recovery procedures involve recreating application configurations, resyncing data sources, and redeploying web experiences
- Organizations can implement disaster recovery strategies by maintaining configuration backups and ensuring source data system recovery capabilities

Recovery Objectives

Amazon Q Business helps meet RPO objectives through continuous data synchronization from source systems and configurable sync schedules (hourly, daily, weekly, or custom intervals). For RTO objectives, the fully managed service provides automatic infrastructure recovery with minimal manual intervention required. Organizations can achieve low RTO by maintaining Infrastructure as Code templates for quick application environment recreation and ensuring rapid data source reconnection. The service's stateless architecture and API-driven configuration enable fast recovery processes, though specific RPO/RTO values depend on data source synchronization frequency and recovery procedures implementation.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/amazonq/latest/qbusiness-ug/quotas-regions.html

FAQ: https://aws.amazon.com/q/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/amazonq/latest/qbusiness-ug/what-is.html

User Guide: https://docs.aws.amazon.com/amazonq/latest/qbusiness-ug/getting-started.html

API Reference: https://docs.aws.amazon.com/amazonq/latest/api-reference/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/q/business/pricing/

Cost Optimization

Amazon Q Business offers several unique cost optimization opportunities including tiered subscription models (Lite at $3/user/month vs Pro at $20/user/month) allowing organizations to match features to user needs. Index capacity optimization through Starter ($0.140/hour) vs Enterprise ($0.264/hour) index types enables right-sizing for different workloads:

- Organizations can optimize connector usage by configuring efficient sync schedules and using incremental rather than full synchronization to reduce processing costs
- The consumption-based anonymous user model provides cost-effective options for public-facing applications
- Additional savings come from reducing manual knowledge management overhead, eliminating custom AI infrastructure costs, and leveraging included connector hours to minimize additional data transfer charges

Savings Considerations

Primary cost savings include eliminating the need for custom AI infrastructure, data science teams, and machine learning model development and maintenance. Organizations save on knowledge management overhead through automated information retrieval and reduced time spent searching for information across multiple systems:

- The subscription model provides predictable costs compared to usage-based alternatives, while the ability to mix Lite and Pro subscriptions optimizes spending based on user requirements
- Additional savings result from reduced training costs through self-service information access, decreased IT support tickets through automated assistance, and improved employee productivity through streamlined workflows
- The fully managed service eliminates operational overhead costs associated with maintaining AI systems, updating models, and ensuring security compliance

Service Name

Amazon Q Developer

Service Overview

Amazon Q Developer is a generative AI-powered conversational assistant built on Amazon Bedrock that helps understand, build, extend, and operate AWS applications. It provides software development assistance through code chat, inline code completions, generating new code, security vulnerability scanning, and code upgrades. Available across IDEs, AWS Management Console, chat applications, and command line interfaces, it accelerates software development by offering real-time coding assistance, automated testing, code reviews, and application transformations. The service continuously updates its capabilities to provide contextually relevant and actionable answers for AWS architecture, resources, best practices, and support.

Key Use Cases

- Code generation and completion: Generate full functions, code snippets, and provide inline suggestions across multiple programming languages in IDEs
- Code security and quality: Automated security vulnerability scanning, code quality reviews, and detection of code smells and anti-patterns
- Application transformation: Automated Java and .NET application upgrades, language version updates, and database migration assistance
- Documentation and testing: Generate comprehensive documentation, unit tests, and improve test coverage automatically
- AWS operations: Diagnose console errors, analyze AWS resources, provide architectural guidance, and troubleshoot operational issues

Features

- **Agentic Coding**: Interactive development assistance through natural language conversations, enabling complex task completion with multi-step reasoning
- **Code Generation and Completion**: Real-time inline code suggestions, full function generation, and intelligent code completion across multiple programming languages
- **Security Scanning**: Automated security vulnerability detection, secrets scanning, and infrastructure security analysis using generative AI and rule-based reasoning
- **Code Reviews**: Automated code quality analysis detecting code smells, anti-patterns, naming violations, and providing on-demand fixes
- **Unit Test Generation**: Automatic generation and placement of unit tests with self-debugging capabilities to improve test coverage

- **Documentation Generation**: Comprehensive documentation creation including READMEs, data flow diagrams, and API documentation
- **Code Transformation**: Automated language upgrades, framework updates, and database migration assistance for Java and .NET applications
- **Multi-Platform Integration**: Available across IDEs (VS Code, JetBrains, Eclipse, Visual Studio), AWS Console, GitHub, GitLab, and chat applications

Value Proposition

Amazon Q Developer offers unique advantages with its deep AWS integration, providing contextual assistance for cloud-native development beyond generic coding tools. Built on Amazon Bedrock with high-quality AWS content, it delivers more complete and actionable answers:

- The service combines multiple development workflows into a single platform - from code generation to security scanning, testing, and documentation
- Key differentiators include agentic capabilities for complex multi-step tasks, real-time AWS resource analysis, automated application transformations, and enterprise-grade security with IP indemnity for Pro users
- Unlike standalone coding assistants, it provides comprehensive AWS architecture guidance, cost optimization recommendations, and operational troubleshooting capabilities

Service Integration

Amazon Q Developer integrates extensively across the AWS ecosystem, powered by Amazon Bedrock for its foundation models. Key integrations include AWS Lambda for serverless development assistance, Amazon SageMaker Studio and Canvas for ML model building, Amazon Redshift for generative SQL capabilities, AWS CloudShell and CLI for command-line assistance, and Amazon OpenSearch Service for operational analytics:

- The service connects with AWS Organizations for enterprise management, AWS Identity Center for workforce authentication, and AWS Backup through architectural recommendations
- It also integrates with AWS CloudWatch for monitoring insights, AWS Systems Manager for operational guidance, and provides native support within the AWS Management Console, mobile app, and documentation websites

Onboarding and Offboarding

Getting started with Amazon Q Developer involves multiple access paths based on user needs. For individual developers, sign up with a Builder ID on the Amazon Q Developer website or install IDE extensions (VS Code, JetBrains, Visual Studio, Eclipse) for free access:

- Organizations use AWS IAM Identity Center for workforce identities and Pro subscriptions
- Access the service through AWS Management Console with required IAM permissions, team chat applications (Slack, Microsoft Teams) with AmazonQDeveloperAccess policy, or command-line interfaces
- The service offers a perpetual Free tier with monthly limits and a Pro tier ($19/month) with expanded capabilities
- Installation is straightforward through marketplace downloads for supported IDEs

Getting Started Guide: https://docs.aws.amazon.com/amazonq/latest/qdeveloper-ug/getting-started-q-dev.html

Data Removal

Offboarding from Amazon Q Developer involves unsubscribing through the Amazon Q console by navigating to the Subscriptions page, selecting users, and choosing the unsubscribe action. Upon cancellation, users immediately lose access to Amazon Q Developer features. Canceling mid-month still incurs the full month's subscription fee without refund. Data stored by Amazon Q Developer includes user questions, responses, and contextual information for generating responses. The service stores data in different AWS regions based on tier and feature usage, with Pro tier content typically stored in the same region as the user's profile.

Backup/Restore and Disaster Recovery

Amazon Q Developer as a managed AWS service does not require traditional backup and disaster recovery as it is a stateless conversational AI service. User interactions and conversation history are temporarily stored for context but are not persistent data requiring backup. The underlying Amazon Bedrock foundation models and AWS infrastructure provide inherent resilience. Organizations using Amazon Q Developer should focus on backing up their own code repositories and development environments rather than the service itself, as Amazon Q Developer generates recommendations based on current context rather than maintaining historical state.

Backup Capabilities

Amazon Q Developer does not have traditional backup capabilities as it is a conversational AI service that operates on current context rather than persistent data storage. The service does not integrate with AWS Backup since it doesn't generate data requiring backup:

- Instead, conversation history is automatically managed with temporary storage for context within sessions
- Users should focus on backing up their own source code, development artifacts, and IDE configurations that Amazon Q Developer helps create or modify

Disaster Recovery

Amazon Q Developer does not directly support traditional disaster recovery scenarios as it is a managed, stateless service built on resilient AWS infrastructure. The service does not integrate with AWS Elastic Disaster Recovery since there are no persistent resources to recover. However, Amazon Q Developer can assist with disaster recovery planning by providing recommendations for resilient design patterns, multi-AZ deployments, and disaster recovery best practices based on the AWS Well-Architected Framework for applications being developed.

Recovery Objectives

Amazon Q Developer helps customers meet RPO and RTO objectives indirectly through development guidance rather than direct recovery capabilities. The service provides recommendations for implementing resilient architectures, disaster recovery patterns, and automated recovery procedures. It can assist in designing multi-region deployments, configuring automated backups for applications, and implementing infrastructure as code for rapid recovery. Amazon Q Developer also helps create disaster recovery testing frameworks and failure mode analysis, enabling faster recovery times through better-designed applications and operational procedures.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/general/latest/gr/amazonqdev.html

FAQ: https://aws.amazon.com/q/developer/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/amazonq/latest/qdeveloper-ug/what-is.html

User Guide: https://docs.aws.amazon.com/amazonq/latest/qdeveloper-ug/getting-started-q-dev.html

API Reference:
https://docs.aws.amazon.com/chatbot/latest/APIReference/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/q/developer/pricing/

Cost Optimization

Amazon Q Developer offers significant cost optimization through its Free tier providing perpetual access to core features with monthly limits, making it accessible for individual developers and small teams. The Pro tier at $19/month per user includes pooled transformation limits at the account level, allowing organizations to share allocations efficiently:

- Transformation capabilities use pay-per-use pricing ($0.003 per line of code) beyond included limits, enabling predictable cost management
- The service eliminates the need for multiple development tools by consolidating coding assistance, security scanning, testing, and documentation generation into a single platform, reducing overall tooling costs and development time

Savings Considerations

Primary cost savings come from accelerated development cycles through AI-assisted coding, reducing time-to-market and developer productivity increases. Organizations save on multiple tool licenses by consolidating coding assistance, security scanning, code review tools, and documentation generation into Amazon Q Developer:

- The service reduces the need for specialized security and code quality consultants through automated scanning and recommendations
- Transformation capabilities save significant migration costs for Java and .NET application upgrades
- Free tier usage for smaller teams and individual developers provides substantial value without ongoing costs, while Pro tier's pooled limits across accounts optimize resource utilization for larger organizations

Service Name

Amazon Redshift

Service Overview

Amazon Redshift is a fully managed, petabyte-scale cloud data warehouse service designed for analyzing structured data using standard SQL and business intelligence tools. It uses massively parallel processing (MPP) architecture with columnar storage technology to deliver fast performance at scale. Amazon Redshift offers both provisioned clusters and serverless options, enabling organizations to run complex analytical workloads without managing infrastructure while providing the performance of traditional data warehousing engines at significantly lower costs.

Key Use Cases

- Business Intelligence and Reporting: Running enterprise BI workloads, generating dashboards, and creating analytical reports for sales, finance, and operational data
- Real-time Analytics: Analyzing streaming data from IoT devices, clickstreams, gaming telemetry, and social media for immediate insights and decision-making
- Data Lake Analytics: Querying data stored in Amazon S3 using Redshift Spectrum without loading data into the warehouse, enabling analysis of huge volumes of structured and semi-structured data
- Machine Learning and Predictive Analytics: Using Redshift ML to create, train, and deploy ML models directly within the data warehouse for fraud detection, risk scoring, and customer churn prediction
- Cross-organizational Data Sharing: Securely sharing live data between different teams, departments, or business units using data sharing capabilities without copying or moving data

Features

- **RA3 Instances with Managed Storage**: Separate compute and storage scaling with high-performance query processing and automatic storage management
- **Concurrency Scaling**: Automatically provisions additional compute capacity to handle varying workload demands without performance degradation
- **Redshift Spectrum**: Query data directly in Amazon S3 using SQL without loading it into Redshift tables
- **Zero-ETL Integrations**: Near real-time data replication from Aurora, RDS, DynamoDB, and enterprise applications without complex ETL pipelines
- **Serverless Option**: Automatically manages compute capacity with AI-driven scaling and optimization without infrastructure management

- **Data Sharing**: Secure sharing of live data between separate Redshift clusters without copying or moving data
- **Redshift ML**: Create, train, and apply machine learning models using SQL commands with SageMaker integration
- **Multi-AZ Deployment**: High availability with automatic failover and data replication across multiple Availability Zones

Value Proposition

Amazon Redshift provides up to 6x better price performance compared to other cloud data warehouses and up to 7x better performance for high concurrency workloads. It offers the flexibility of both provisioned and serverless deployment models, allowing customers to optimize for their specific needs:

- The service eliminates the complexity of traditional data warehousing with automatic scaling, built-in security features including end-to-end encryption, and seamless integration with the broader AWS ecosystem
- Customers benefit from no upfront costs, pay-as-you-go pricing, and the ability to scale from gigabytes to petabytes without performance degradation

Service Integration

Amazon Redshift integrates natively with numerous AWS services including Amazon S3 for data lake analytics through Redshift Spectrum, Amazon Aurora and RDS for zero-ETL data replication, Amazon SageMaker for machine learning capabilities, Amazon Kinesis for streaming data ingestion, AWS Glue for ETL processes, Amazon QuickSight for business intelligence dashboards, AWS IAM for security and access control, Amazon CloudWatch for monitoring, AWS Backup for centralized backup management, and Amazon VPC for network isolation. It also supports integration with AWS Lake Formation for data cataloging and governance.

Onboarding and Offboarding

Customers can start with Amazon Redshift by creating either a provisioned cluster or serverless workgroup through the AWS Management Console, CLI, or API. The Getting Started experience includes sample datasets and query examples to familiarize users with the service:

- For provisioned clusters, customers choose node types and cluster size based on their requirements
- For serverless, they simply specify base capacity settings
- The service provides automated setup of VPC configurations, security groups, and parameter groups with sensible defaults, allowing users to begin querying data within minutes of creation

Getting Started Guide: https://docs.aws.amazon.com/redshift/latest/gsg/

Data Removal

To offboard from Amazon Redshift, customers can delete their clusters or serverless workgroups, which removes all associated compute resources. Manual snapshots must be explicitly deleted to remove data permanently, while automated snapshots are deleted when the cluster is terminated. For complete data removal, customers should delete all snapshots, remove any cross-region snapshot copies, and clean up associated resources like parameter groups and subnet groups to ensure no residual data or costs remain.

Backup/Restore and Disaster Recovery

Amazon Redshift provides automated and manual snapshot capabilities with configurable retention periods. Snapshots are stored in Amazon S3 and can be copied across regions for disaster recovery. The service supports point-in-time recovery through continuous automated snapshots taken every 8 hours or after 5 GB of data changes. Multi-AZ deployments offer automatic failover with sub-minute recovery times and synchronous data replication for high availability scenarios.

Backup Capabilities

Amazon Redshift includes comprehensive backup capabilities with automated snapshots taken periodically and manual snapshots created on-demand. The service integrates with AWS Backup for centralized backup management and policy-based backup scheduling. Backup retention periods are configurable from 1 to 35 days for automated snapshots, while manual snapshots can be retained indefinitely or for specified periods up to 3,653 days.

Disaster Recovery

Amazon Redshift supports disaster recovery through cross-region snapshot copying, Multi-AZ deployments for automatic failover, and integration with AWS Backup for centralized disaster recovery management. While not directly integrated with AWS Elastic Disaster Recovery, customers can implement disaster recovery strategies using snapshot restoration, data sharing across regions, and architectural patterns that leverage multiple AWS regions for business continuity.

Recovery Objectives

Amazon Redshift helps meet RPO objectives through automated snapshots every 8 hours or after 5 GB of data changes, providing recovery points with minimal data loss. RTO objectives are supported through Multi-AZ deployments with sub-minute automatic failover, elastic resize for quick capacity changes, and snapshot restoration capabilities. Cross-region snapshot copying enables disaster recovery with RPO and RTO aligned to business requirements and snapshot frequency configurations.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/redshift/latest/mgmt/amazon-redshift-limits.html

FAQ: https://aws.amazon.com/redshift/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/redshift/

User Guide: https://docs.aws.amazon.com/redshift/latest/mgmt/

API Reference: https://docs.aws.amazon.com/redshift/latest/APIReference/

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/redshift/pricing/

Cost Optimization

Amazon Redshift offers multiple cost optimization strategies including Reserved Instances with up to 75% savings over on-demand pricing, Serverless Reservations providing up to 24% discounts on committed RPU usage, and pause/resume functionality for development environments. Customers can optimize costs through proper data compression, efficient table design, and using Redshift Spectrum for infrequently accessed data. Concurrency Scaling provides one hour free per day, and managed storage in RA3 instances prevents over-provisioning by scaling storage independently of compute resources.

Savings Considerations

Primary cost savings come from choosing appropriate pricing models (Reserved Instances vs on-demand), rightsizing clusters based on actual workload requirements, and leveraging Serverless for variable workloads to avoid idle capacity costs. Additional savings are achieved through data lifecycle management using Redshift Spectrum for cold data, optimizing query performance to reduce compute usage, and using pause/resume for non-production environments. Cross-region data transfer costs should be considered when implementing multi-region architectures, and backup storage costs can be managed by setting appropriate retention periods for manual snapshots.

Service Name

Amazon Rekognition

Service Overview

Amazon Rekognition is a cloud-based image and video analysis service that enables developers to add advanced computer vision capabilities to their applications without requiring machine learning expertise. It uses deep learning technology to detect and analyze objects, scenes, concepts, faces, text, celebrities, and inappropriate content in billions of images and videos daily. The service offers both synchronous image analysis through Amazon Rekognition Image and asynchronous video analysis through Amazon Rekognition Video, supporting both stored videos and real-time streaming analysis.

Key Use Cases

- Searchable media libraries: Enable content discovery and organization through automated tagging and metadata extraction
- Identity verification and face-based authentication: Verify user identities through face matching and liveness detection
- Content moderation: Automatically detect and filter inappropriate, explicit, or unsafe content in images and videos
- Security and surveillance: Monitor premises through real-time detection of people, objects, and activities
- Media analysis and workflow automation: Streamline video production with shot detection, segment identification, and frame-accurate analysis
- Personal protective equipment (PPE) detection: Ensure workplace safety compliance by detecting safety equipment usage
- Text extraction and document processing: Extract and analyze text from images for document digitization and processing

Features

- **Object and Scene Detection**: Detects thousands of objects, scenes, concepts, and activities with confidence scores and bounding boxes
- **Facial Analysis and Recognition**: Analyzes facial attributes, emotions, landmarks, and enables face comparison and search capabilities
- **Text Detection and Extraction**: Detects and extracts text from images and videos with word-level confidence scores
- **Celebrity Recognition**: Identifies thousands of celebrities with additional metadata including names and IMDb IDs
- **Content Moderation**: Detects explicit, suggestive, or violent content to enable automated content filtering

- **Custom Labels**: Creates custom models to detect specific objects, scenes, or concepts unique to business needs
- **Video Streaming Events**: Processes real-time video streams for low-latency object detection including people, pets, and packages
- **PPE Detection**: Identifies personal protective equipment including face covers, head covers, and hand covers
- **Media Analysis**: Provides frame-accurate segment detection including shot changes, black frames, and color bars
- **Face Liveness Detection**: Verifies presence of live users during authentication to prevent spoofing attacks

Value Proposition

Amazon Rekognition eliminates the complexity and cost of building custom computer vision solutions by providing pre-trained, highly accurate deep learning models accessible through simple APIs. It offers unmatched scalability to process billions of images and videos, integrates seamlessly with other AWS services, and requires no machine learning expertise:

- The service provides enterprise-grade security and compliance including HIPAA eligibility, while offering cost-effective pay-per-use pricing with generous free tiers
- Customers benefit from continuous model improvements, responsible AI practices, and the ability to customize models for specific business needs through Custom Labels

Service Integration

Amazon Rekognition integrates deeply with Amazon S3 for storing and processing images and videos, AWS Lambda for serverless processing workflows and event-driven automation, Amazon SNS for notifications when analysis jobs complete, Amazon SQS for decoupling processing workflows, Amazon Kinesis Video Streams for real-time video analysis, Amazon CloudWatch for monitoring and metrics, AWS CloudTrail for API logging and auditing, Amazon DynamoDB for storing analysis results and metadata, and Amazon Augmented AI (A2I) for human review workflows. It also supports integration with Amazon EventBridge for event-driven architectures and IAM for fine-grained access control.

Onboarding and Offboarding

Getting started requires creating an AWS account and setting up appropriate IAM permissions for Amazon Rekognition access. Customers then install and configure the AWS CLI and SDKs for their preferred programming languages:

- The service offers a web-based console for testing capabilities, comprehensive developer guides with code examples in multiple languages, and step-by-step tutorials for common use cases

- Amazon provides ready-to-use CloudFormation templates, sample applications, and integration patterns with other AWS services to accelerate adoption
- The free tier allows customers to process 1,000 images and 60 minutes of video monthly at no cost

Getting Started Guide: https://docs.aws.amazon.com/rekognition/latest/dg/getting-started.html

Data Removal

Customers retain full control and ownership of their content processed by Amazon Rekognition. Image and video inputs can be deleted at any time through the console or APIs. Face collections and associated metadata can be removed using DeleteCollection and DeleteFaces operations. For complete offboarding, customers should delete all face collections, remove stored analysis results, and revoke IAM permissions. AWS automatically deletes temporary processing data, and customers can use AWS CloudTrail to audit all data access and deletion activities.

Backup/Restore and Disaster Recovery

Amazon Rekognition is a managed service where AWS handles infrastructure backup and disaster recovery. Face collections and user vectors are automatically replicated across multiple Availability Zones for high availability. Customers are responsible for backing up their own analysis results, face collection metadata, and application configurations. Analysis results stored in other AWS services like DynamoDB or S3 can leverage those services' backup capabilities including AWS Backup integration.

Backup Capabilities

Amazon Rekognition does not directly integrate with AWS Backup as it is a fully managed API service. However, customers can backup face collections by exporting face vectors and metadata, and analysis results stored in integrated services like DynamoDB or S3 can utilize AWS Backup. The service maintains automatic replication of face collections for high availability, and customers should implement backup strategies for their analysis workflows and results.

Disaster Recovery

Amazon Rekognition provides built-in disaster recovery through multi-AZ replication of face collections and service infrastructure. The service does not directly integrate with AWS Elastic Disaster Recovery, but customers can implement disaster recovery for their applications using Rekognition by replicating face collections across regions, backing up analysis results, and designing multi-region architectures. AWS manages the underlying service resilience and availability.

Recovery Objectives

Amazon Rekognition supports low RPO and RTO objectives through its multi-AZ architecture and high availability design. Face collections are automatically replicated, enabling near-zero RPO for face data. For comprehensive disaster recovery, customers should replicate face collections across regions, backup analysis results, and design applications with failover capabilities. The service's API-based architecture enables rapid recovery and regional failover when properly architected.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/rekognition/latest/dg/limits.html

FAQ: https://aws.amazon.com/rekognition/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/rekognition/latest/dg/what-is.html

User Guide: https://docs.aws.amazon.com/rekognition/latest/dg/getting-started.html

API Reference: https://docs.aws.amazon.com/rekognition/latest/APIReference/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/rekognition/pricing/

Cost Optimization

Amazon Rekognition offers several cost optimization opportunities including tiered pricing that decreases cost per image with higher volumes, free tier providing 1,000 images and 60 minutes of video monthly, ability to set confidence thresholds to process only relevant content, custom moderation adapters to reduce false positives, bulk analysis capabilities for batch processing efficiency, and regional pricing variations. Customers can optimize by choosing appropriate AWS regions, implementing intelligent filtering, using batch processing, and leveraging Custom Labels for specific use cases to avoid processing irrelevant content.

Savings Considerations

Primary cost savings come from leveraging the generous free tier limits, implementing volume-based tiered pricing benefits, using batch processing for large datasets, setting optimal confidence thresholds to reduce unnecessary processing, utilizing Custom Labels

to focus on business-specific objects, choosing cost-effective regions, integrating with AWS Lambda for event-driven processing to avoid continuous polling, and implementing intelligent pre-filtering to process only relevant media. Organizations can achieve significant savings by designing efficient workflows, avoiding redundant analysis, and taking advantage of the pay-per-use model with no upfront commitments or minimum fees.

Service Name

Amazon Relational Database Service (RDS)

Service Overview

Amazon RDS is a fully managed relational database service that simplifies the setup, operation, and scaling of relational databases in the AWS Cloud. It supports multiple database engines including MySQL, PostgreSQL, MariaDB, Oracle Database, Microsoft SQL Server, and IBM Db2. RDS automates essential database administration tasks such as hardware provisioning, database setup, patching, backups, and recovery, allowing users to focus on their applications. The service provides cost-efficient, resizable capacity while maintaining compatibility with existing database applications and tools.

Key Use Cases

- Web and mobile applications: Supporting transactional workloads with high availability and automated scaling for consumer-facing applications
- Enterprise applications: Running business-critical applications requiring ACID compliance, complex queries, and advanced analytics capabilities
- E-commerce platforms: Managing product catalogs, customer data, and transaction processing with built-in security and backup features
- Content management systems: Storing and managing structured content with read replicas for improved performance
- Analytics and reporting: Supporting OLTP workloads with read replicas for business intelligence and reporting without impacting primary operations

Features

- **Multi-AZ Deployments**: Provides synchronous replication to a standby instance in a different Availability Zone for high availability and automatic failover
- **Read Replicas**: Creates read-only copies of databases to scale read workloads and improve application performance
- **Automated Backups**: Performs daily automated backups with point-in-time recovery capability up to 35 days
- **Security Features**: Provides encryption at rest and in transit, VPC isolation, IAM integration, and database authentication
- **Performance Insights**: Offers database performance monitoring and analysis tools for query optimization
- **Blue/Green Deployments**: Enables zero-downtime database updates and deployments through managed environment switching

Value Proposition

Amazon RDS eliminates the operational complexity of database management while providing enterprise-grade reliability and security. Key advantages include 69% cost savings with Reserved Instances, automatic patching and updates, built-in monitoring and alerting, and seamless integration with AWS services:

- RDS reduces administrative overhead by up to 43% compared to self-managed databases, provides 99.95% availability SLA with Multi-AZ deployments, and offers flexible scaling without downtime
- The service ensures compatibility with existing applications while providing advanced features like automated failover, read replicas, and comprehensive backup solutions

Service Integration

RDS integrates deeply with AWS Backup for centralized backup management across multiple services, AWS CloudWatch for monitoring and alerting, and AWS IAM for access control and authentication. Key integrations include VPC for network isolation, AWS KMS for encryption key management, Amazon S3 for backup storage, and AWS Secrets Manager for credential management:

- RDS also works with AWS DMS for database migration, Amazon ElastiCache for caching layers, AWS Lambda for serverless applications, and Amazon CloudTrail for audit logging
- Performance optimization integrations include Amazon CloudWatch Performance Insights and AWS Systems Manager for parameter management

Onboarding and Offboarding

Getting started with RDS involves creating an AWS account, selecting a database engine through the RDS console, and using the Easy Create option for simplified setup. The process includes choosing instance class, storage type, and basic networking configuration:

- AWS provides step-by-step tutorials for each supported database engine, guided workflows in the console, and CLI automation options
- Initial setup typically takes 10-20 minutes with automated provisioning
- Best practices include configuring VPC security groups, setting up parameter groups for optimization, and establishing backup schedules
- Free Tier eligibility provides 750 hours monthly for 12 months to evaluate the service

Getting Started Guide:
https://docs.aws.amazon.com/AmazonRDS/latest/gettingstartedguide/what-is-rds.html

Data Removal

RDS data removal involves deleting DB instances through the console, CLI, or API with options to create final snapshots before deletion. Automated backups are deleted when instances are removed unless specifically retained. Manual snapshots persist until explicitly deleted and can be shared across accounts or regions. The deletion process permanently removes all data and affects related resources like read replicas. Organizations can implement data retention policies and use AWS Organizations for centralized control over deletion permissions across multiple accounts.

Backup/Restore and Disaster Recovery

RDS provides automated daily backups with transaction log captures every 5 minutes for point-in-time recovery up to 35 days. Manual snapshots offer indefinite retention and can be copied across regions. Multi-AZ deployments provide synchronous replication for high availability and automatic failover. Cross-region automated backups enable disaster recovery with asynchronous replication. AWS Backup integration allows centralized backup management with lifecycle policies and compliance features.

Backup Capabilities

RDS includes native automated backup capabilities with configurable retention periods up to 35 days and supports integration with AWS Backup for centralized management. Automated backups capture full daily snapshots and continuous transaction logs for granular point-in-time recovery:

- Manual snapshots provide user-initiated backups with indefinite retention
- Backup encryption, cross-region copying, and account sharing capabilities are included
- Performance impact is minimized by taking backups from standby instances in Multi-AZ deployments

Disaster Recovery

RDS supports comprehensive disaster recovery through Multi-AZ deployments for same-region protection and cross-region read replicas for geographic redundancy. Cross-region automated backups provide asynchronous disaster recovery with managed replication:

- The service does not directly integrate with AWS Elastic Disaster Recovery but provides native disaster recovery capabilities through automated failover, backup restoration, and replica promotion
- Recovery strategies include active-passive configurations with standby databases and active-active setups using read replicas

Recovery Objectives

RDS enables low Recovery Point Objectives (RPO) of minutes through Multi-AZ synchronous replication and 5-minute transaction log intervals. Recovery Time Objectives (RTO) range from 1-2 minutes for automatic Multi-AZ failover to several hours for cross-region disaster recovery. Point-in-time recovery allows restoration to any second within the backup retention period. Read replicas can be quickly promoted to primary instances for faster recovery. Organizations can design architectures meeting specific RPO/RTO requirements using combinations of Multi-AZ, read replicas, and backup strategies.

Service Constraints

Service Quotas:
https://docs.aws.amazon.com/AmazonRDS/latest/UserGuide/CHAP_Limits.html

FAQ: https://aws.amazon.com/rds/faqs/

Technical Requirements

Service Documentation:
https://docs.aws.amazon.com/AmazonRDS/latest/UserGuide/Welcome.html

User Guide: https://docs.aws.amazon.com/AmazonRDS/latest/UserGuide/Welcome.html

API Reference:
https://docs.aws.amazon.com/AmazonRDS/latest/UserGuide/ProgrammingGuide.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/rds/pricing/

Cost Optimization

RDS offers Reserved Instances with up to 69% savings over On-Demand pricing through 1-year or 3-year commitments with flexible payment options. Instance size flexibility allows changing instance types within the same family:

- Aurora Serverless provides automatic scaling for variable workloads
- Storage optimization includes choosing appropriate storage types (gp2, gp3, io1) based on performance requirements
- AWS Compute Optimizer provides right-sizing recommendations based on actual usage patterns
- Graviton-based instances offer better price-performance ratios
- Automated scaling prevents over-provisioning while maintaining performance

Savings Considerations

Primary cost savings come from Reserved Instance commitments offering significant discounts for predictable workloads, with All Upfront payments providing maximum savings. Right-sizing instances based on actual CPU and memory utilization can reduce costs by 20-40%:

- Storage optimization through appropriate type selection and compression reduces storage costs
- Read replicas can reduce primary instance load, allowing smaller primary instances
- Automated backup lifecycle management prevents unnecessary backup storage costs
- Multi-engine cost comparison helps select the most cost-effective database engine
- Regular monitoring and AWS Cost Explorer usage optimize ongoing expenses

Service Name

Amazon Route 53

Service Overview

Amazon Route 53 is a highly available and scalable Domain Name System (DNS) web service that performs four main functions: domain registration, DNS routing, health checking, and DNS resolution. Route 53 helps users register domain names, translates friendly domain names into IP addresses, sends automated requests to verify application availability, and forwards DNS queries between VPCs and networks. The service uses a global network of authoritative DNS servers to reduce latency and provides comprehensive DNS management capabilities for both public and private domains, making it an essential component for internet-facing applications and hybrid cloud architectures.

Key Use Cases

- Domain Registration: Register and manage domain names with various top-level domains (TLDs) and privacy protection options
- DNS Traffic Routing: Route internet traffic to resources using advanced routing policies like latency-based, geolocation, weighted, and failover routing
- Health Monitoring and Failover: Monitor endpoint health and automatically route traffic away from unhealthy resources to maintain application availability
- Hybrid DNS Resolution: Enable DNS resolution between on-premises networks and AWS VPCs using Route 53 Resolver for seamless hybrid cloud connectivity
- Global Traffic Management: Distribute traffic across multiple regions and resources based on performance, location, or business requirements
- Private DNS Management: Manage internal DNS resolution within VPCs using private hosted zones for secure internal communication

Features

- **DNS Hosting**: Public and private hosted zones for authoritative DNS management with global anycast network
- **Advanced Routing Policies**: Latency-based, geolocation, geoproximity, weighted, failover, IP-based, and multivalue routing
- **Health Checks**: Automated monitoring of endpoints with DNS failover and CloudWatch integration
- **Route 53 Resolver**: Recursive DNS resolution between VPCs and on-premises networks with conditional forwarding
- **Domain Registration**: Register and manage domains with privacy protection and auto-renewal options

- **DNSSEC Support**: DNS Security Extensions for authentication and integrity of DNS responses
- **Traffic Flow**: Visual editor for complex routing policies and global traffic management
- **Resolver DNS Firewall**: Filter and block malicious DNS queries with customizable rules
- **Route 53 Profiles**: Centralized DNS configuration management across multiple VPCs and accounts
- **Global Resolver**: Secure anycast DNS resolution with encrypted queries and centralized policies

Value Proposition

Route 53 provides unmatched reliability with 100% availability SLA, leveraging AWS's global infrastructure of edge locations for low-latency DNS responses. The service offers comprehensive DNS management combining domain registration, authoritative DNS, health monitoring, and hybrid DNS resolution in a single integrated platform:

- Key differentiators include advanced routing capabilities for intelligent traffic distribution, seamless integration with AWS services through alias records, and enterprise-grade security features including DNSSEC and DNS Firewall
- Route 53's pay-as-you-go pricing model eliminates upfront costs while providing predictable scaling
- The service's global anycast network ensures optimal performance worldwide, while features like Accelerated Recovery provide business continuity during regional disruptions

Service Integration

Route 53 integrates deeply with core AWS services through alias records and direct service integration. Primary integrations include Amazon CloudFront for global content delivery, Elastic Load Balancing (ALB, NLB, CLB) for application traffic distribution, and Amazon S3 for static website hosting:

- The service connects with Amazon EC2 instances, AWS Elastic Beanstalk environments, and Amazon API Gateway endpoints for application routing
- Route 53 Resolver integrates with Amazon VPC for private DNS resolution and AWS Direct Connect for hybrid connectivity
- Health checks integrate with Amazon CloudWatch for monitoring and Amazon SNS for notifications
- The service leverages AWS KMS for DNSSEC key management and AWS CloudTrail for audit logging
- Integration with Amazon WorkMail enables email routing, while AWS Identity and Access Management provides granular access control

Onboarding and Offboarding

Getting started with Route 53 requires an AWS account and follows a straightforward process. Begin by signing up for AWS and setting up IAM permissions for Route 53 access:

- For domain registration, search for available domains and complete the registration process with contact information
- For DNS hosting, create a hosted zone for your domain and configure DNS records
- The service provides step-by-step tutorials for common scenarios like routing traffic to S3 websites or CloudFront distributions
- Route 53 offers a management console, APIs, CLI, and SDK access for different implementation approaches
- New users can leverage the getting started guide that walks through registering a domain, creating S3 bucket websites, and configuring DNS routing
- The service includes built-in validation tools and health checks to verify proper configuration

Getting Started Guide:
https://docs.aws.amazon.com/Route53/latest/DeveloperGuide/getting-started.html

Data Removal

Route 53 data removal involves deleting hosted zones, records, health checks, and domain configurations. To delete a hosted zone, ensure only NS and SOA records remain, then delete through the console or API. Domain registrations require transfer to another registrar before deletion from Route 53. Health checks, traffic policies, and resolver configurations can be deleted individually. For complete offboarding, remove all DNS records, disable DNSSEC signing if enabled, delete resolver endpoints and rules, and cancel domain registrations. The service provides APIs for programmatic deletion and maintains audit logs in CloudTrail for compliance requirements.

Backup/Restore and Disaster Recovery

Route 53 provides built-in disaster recovery through its globally distributed architecture and data plane redundancy. The service maintains automatic replication across multiple edge locations worldwide. Route 53 Accelerated Recovery offers 60-minute RTO for DNS operations during regional disruptions by failing over the control plane to secondary regions. Customers can implement backup strategies by exporting DNS records using APIs and storing configurations in version control systems. The service's anycast network ensures continued DNS resolution even during localized outages, while health checks enable automatic failover to healthy endpoints.

Backup Capabilities

Route 53 does not integrate with AWS Backup service as it manages DNS data differently from traditional backup scenarios. However, the service provides inherent backup capabilities through its globally distributed architecture that automatically replicates DNS data across multiple locations:

- Customers can export DNS records and configurations using APIs, CLI, or third-party tools for external backup storage
- The service maintains configuration history through CloudTrail logging and supports programmatic recreation of DNS resources for disaster recovery scenarios

Disaster Recovery

Route 53 does not directly integrate with AWS Elastic Disaster Recovery, but it serves as a critical component in disaster recovery architectures. The service enables DNS failover between primary and secondary sites, supports health checks for automatic traffic redirection during outages, and provides geographically distributed DNS resolution:

- Route 53's Accelerated Recovery feature offers built-in control plane failover capabilities
- The service's global infrastructure and anycast network provide inherent disaster recovery benefits by maintaining DNS availability across multiple regions and edge locations

Recovery Objectives

Route 53 helps customers achieve aggressive RPO and RTO objectives through its globally distributed architecture and automated failover capabilities. Health checks enable sub-minute detection of endpoint failures, while DNS failover can redirect traffic within seconds of failure detection. The service's global anycast network ensures continued DNS resolution with minimal impact during localized outages. Route 53 Accelerated Recovery provides 60-minute RTO for DNS management operations during regional disruptions. Customers can configure health check intervals as low as 10 seconds and implement multi-layered failover strategies for mission-critical applications.

Service Constraints

Service Quotas:
https://docs.aws.amazon.com/Route53/latest/DeveloperGuide/DNSLimitations.html

FAQ: https://aws.amazon.com/route53/faqs/

Technical Requirements

Service Documentation:
https://docs.aws.amazon.com/Route53/latest/DeveloperGuide/Welcome.html

User Guide:
https://docs.aws.amazon.com/Route53/latest/DeveloperGuide/Welcome.html

API Reference:
https://docs.aws.amazon.com/Route53/latest/APIReference/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/route53/pricing/

Cost Optimization

Route 53 offers several cost optimization strategies including using alias records instead of CNAME records to reduce query costs, optimizing TTL values to balance responsiveness and query volume, and implementing appropriate health check intervals to minimize unnecessary checks. Customers can leverage Route 53's pay-per-query pricing model by consolidating DNS zones and using wildcard records where appropriate:

- The service's integration with AWS services through alias records eliminates additional query charges for AWS resource routing
- Geographic routing can reduce costs by directing traffic to the most cost-effective regions
- Customers should monitor query volumes and adjust routing policies to optimize both performance and costs while leveraging Route 53 Profiles for centralized management across multiple accounts

Savings Considerations

Primary cost savings opportunities in Route 53 include consolidating multiple DNS providers into a single platform, reducing operational overhead through automated health checks and failover, and eliminating separate DNS hosting fees through AWS service integration. The service's global infrastructure reduces the need for multiple regional DNS providers while maintaining performance:

- Alias records provide cost-free routing to AWS resources compared to traditional CNAME records
- Automated DNS management reduces manual operational costs and potential downtime expenses
- Route 53's integration with other AWS services enables consolidated billing and potential volume discounts
- The pay-as-you-go model eliminates upfront DNS infrastructure investments while providing predictable scaling costs based on actual usage patterns

Service Name

Amazon SageMaker

Service Overview

Amazon SageMaker is a fully managed machine learning service that provides end-to-end ML capabilities for data scientists and developers. It enables users to quickly build, train, and deploy high-quality machine learning models into production-ready hosted environments. The platform offers a comprehensive suite of tools including managed infrastructure, integrated development environments, automated ML capabilities, and deployment options that eliminate the need to manage underlying servers and infrastructure. SageMaker provides both code-based and no-code/low-code options for machine learning workflows, making it accessible to users with varying technical expertise while offering enterprise-grade security, governance, and scalability.

Key Use Cases

- Predictive Analytics: Revenue forecasting, demand planning, customer churn prediction, and business intelligence applications
- Computer Vision: Image classification, object detection, medical image analysis, quality control, and autonomous vehicle development
- Natural Language Processing: Sentiment analysis, text classification, chatbots, document processing, and language translation
- Fraud Detection: Credit card fraud detection, anomaly detection in financial transactions, and risk assessment
- Recommendation Systems: Personalized product recommendations, content recommendations, and targeted marketing campaigns
- Time Series Forecasting: Inventory management, financial forecasting, IoT sensor data analysis, and supply chain optimization
- Foundation Model Development: Training and fine-tuning large language models, generative AI applications, and custom foundation models

Features

- **SageMaker Studio**: Fully integrated development environment with JupyterLab, Code Editor, and RStudio for end-to-end ML development
- **SageMaker Canvas**: No-code visual interface for building ML models with automated data preparation and model training
- **SageMaker Autopilot**: Automated machine learning that automatically builds, trains, and tunes models with full visibility
- **SageMaker Training**: Managed model training with distributed training, hyperparameter tuning, and spot instance support

- **SageMaker HyperPod**: Purpose-built infrastructure for distributed training of foundation models at scale
- **SageMaker Inference**: Multiple deployment options including real-time, batch, asynchronous, and serverless inference
- **SageMaker JumpStart**: Pre-trained models, algorithms, and solution templates for common ML use cases
- **SageMaker Feature Store**: Centralized repository for storing, sharing, and managing ML features across teams
- **SageMaker Pipelines**: Purpose-built CI/CD service for creating end-to-end ML workflows
- **SageMaker Clarify**: Bias detection and model explainability tools for responsible AI

Value Proposition

Amazon SageMaker offers unmatched flexibility with support for any machine learning framework, from scikit-learn to custom deep learning models. It provides the broadest selection of ML tools in a single platform, eliminating the need to stitch together multiple services:

- Key differentiators include managed Spot training that reduces costs by up to 90%, purpose-built infrastructure for foundation model training, integrated data labeling and preparation tools, and enterprise-grade governance features
- The platform scales automatically from experimentation to production, offers both code and no-code options for different user personas, and provides deep integration with the broader AWS ecosystem
- Additionally, SageMaker's pay-as-you-go pricing model and Savings Plans can reduce total cost of ownership by at least 54% compared to self-managed solutions

Service Integration

SageMaker integrates deeply with core AWS data and storage services including Amazon S3 for data lake storage and model artifacts, Amazon Redshift for data warehousing and analytics, DynamoDB for NoSQL data storage with zero-ETL integrations, and AWS Glue for data cataloging and ETL operations. Compute integrations include Amazon EC2 for custom training environments, AWS Batch for large-scale batch processing, and Amazon EMR for big data processing:

- Security and governance integrations encompass AWS IAM for access control, AWS CloudTrail for auditing, AWS KMS for encryption, and VPC for network isolation
- Additional integrations include Amazon Bedrock for foundation models, Amazon Athena for SQL analytics, AWS Lambda for serverless compute, Amazon API Gateway for model serving, and Amazon QuickSight for business intelligence and visualization

Onboarding and Offboarding

Customers can get started with SageMaker through quick setup or custom setup options. The quick setup provides the fastest onboarding for individual users with default configurations, while custom setup offers advanced configuration for enterprise ML administrators:

- SageMaker offers comprehensive resources including sample notebooks, tutorials, and documentation
- New users can access SageMaker Studio Lab without an AWS account for experimentation
- The platform provides guided workflows through SageMaker Canvas for no-code users and comprehensive SDKs and APIs for developers
- Getting started resources include hands-on tutorials, pre-built examples in JumpStart, and integration with popular ML frameworks
- The onboarding process includes domain creation, user profile setup, and access to various development environments based on user preferences and organizational requirements

Getting Started Guide: https://docs.aws.amazon.com/sagemaker/latest/dg/gs.html

Data Removal

Data removal during offboarding involves deleting SageMaker domains, user profiles, and associated resources. Customers must manually delete model artifacts stored in S3, training data, and Feature Store entries. The offboarding process includes stopping active notebook instances, deleting endpoints and model deployments, removing training jobs history, and cleaning up SageMaker projects and pipelines. Organizations should also revoke IAM permissions, remove VPC configurations, and delete any custom Docker images stored in ECR. Complete data removal requires attention to distributed components across multiple AWS services that SageMaker integrates with.

Backup/Restore and Disaster Recovery

SageMaker provides backup capabilities through integration with AWS services rather than native backup features. Model artifacts and training data stored in S3 benefit from S3's cross-region replication and versioning capabilities. Training jobs can be configured for checkpointing to enable recovery from interruptions. SageMaker notebooks support version control integration. For disaster recovery, customers can implement multi-region strategies using S3 cross-region replication, recreate training environments in alternate regions, and deploy models across multiple availability zones. The platform supports event-driven architecture patterns for implementing robust backup and recovery solutions using AWS services like EventBridge, Lambda, and DynamoDB.

Backup Capabilities

SageMaker does not directly integrate with AWS Backup service. Instead, backup capabilities depend on the underlying storage services used:

- S3-stored model artifacts and datasets can be backed up through S3's native features including cross-region replication, versioning, and lifecycle policies
- Training checkpoints provide recovery capabilities for long-running training jobs
- Feature Store data can be backed up through its underlying storage mechanisms
- For comprehensive backup strategies, customers must implement solutions using AWS services like Lambda functions and EventBridge to capture metadata and coordinate backup processes across the SageMaker ecosystem components

Disaster Recovery

SageMaker does not natively integrate with AWS Elastic Disaster Recovery service. Disaster recovery implementations require custom architectures using multiple AWS regions, S3 cross-region replication for data, and Infrastructure as Code templates for rapid environment recreation:

- Organizations can implement pilot light or warm standby strategies by maintaining model artifacts and training data in secondary regions
- The distributed nature of SageMaker components requires comprehensive disaster recovery planning that includes data replication, environment recreation capabilities, and traffic routing strategies using Route 53 and Application Load Balancers for inference endpoints

Recovery Objectives

SageMaker supports varying RPO and RTO objectives through flexible architecture options. For low RPO requirements, customers can implement real-time data replication to S3, use checkpointing for training jobs, and maintain model artifacts across regions. For improved RTO, strategies include pre-provisioned inference endpoints, automated failover using Route 53, and Infrastructure as Code for rapid environment recreation. Multi-AZ deployments provide high availability within regions. The platform's serverless inference option offers automatic scaling and reduced recovery complexity, while managed spot training includes automatic recovery from instance interruptions.

Service Constraints

Service Quotas:
https://docs.aws.amazon.com/general/latest/gr/sagemaker.html#limits_sagemaker

FAQ: https://aws.amazon.com/sagemaker/ai/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/sagemaker/latest/dg/whatis.html

User Guide: https://docs.aws.amazon.com/sagemaker/latest/dg/whatis.html

API Reference: https://docs.aws.amazon.com/sagemaker/latest/APIReference/

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/sagemaker/pricing/

Cost Optimization

SageMaker offers unique cost optimization through managed Spot training that reduces training costs by up to 90% with automatic fault tolerance and checkpoint recovery. SageMaker Savings Plans provide up to 64% cost reduction for consistent usage patterns:

- The platform supports rightsizing through instance recommendations and usage analytics
- Serverless inference eliminates costs during idle periods with automatic scaling to zero
- Multi-model endpoints allow hosting multiple models on single infrastructure reducing deployment costs
- SageMaker Canvas offers consumption-based pricing for no-code ML
- Training job optimization includes automatic scaling, distributed training capabilities, and smart data loading
- Cost optimization tools include automatic stopping of idle resources, spot instance integration, and detailed cost tracking through AWS Cost Explorer integration

Savings Considerations

Primary cost savings opportunities include using Spot instances for non-time-critical training workloads, implementing SageMaker Savings Plans for predictable usage, and leveraging serverless inference for variable workloads. Multi-model endpoints significantly reduce hosting costs by sharing infrastructure across models:

- Automated resource management prevents idle resource charges through features like automatic scaling and notebook instance lifecycle management
- Data optimization strategies include using S3 Intelligent Tiering, optimizing data formats for faster processing, and implementing efficient feature engineering

- Training optimization includes hyperparameter tuning to reduce training time, distributed training for faster completion, and using built-in algorithms optimized for performance
- Regular cost monitoring and rightsizing recommendations help maintain optimal cost efficiency throughout the ML lifecycle

Service Name

Amazon SageMaker AI

Service Overview

Amazon SageMaker AI is a fully managed machine learning platform that provides end-to-end ML capabilities for building, training, and deploying machine learning models at scale. It offers a comprehensive suite of tools including notebooks, debuggers, profilers, pipelines, MLOps, and governance features within an integrated development environment. The platform supports both traditional machine learning workflows and generative AI development, enabling data scientists and ML engineers to develop foundation models from scratch, customize pre-trained models, and deploy them for inference with enterprise-grade controls and security.

Key Use Cases

- Foundation model development: Building and training large-scale foundation models and generative AI applications using purpose-built infrastructure like SageMaker HyperPod
- Automated machine learning: Creating ML models without extensive coding using SageMaker Canvas for visual model building and SageMaker Autopilot for automated model development
- Enterprise MLOps: Implementing end-to-end ML workflows with governance, bias detection, model monitoring, and automated deployment pipelines for production-scale ML operations

Features

- **SageMaker Studio**: Fully integrated development environment for machine learning with JupyterLab, Code Editor, and collaborative features
- **SageMaker Autopilot**: Automated ML service that builds, trains, and tunes models with full visibility into the process
- **SageMaker Canvas**: Visual, no-code interface for building ML models without requiring programming skills
- **SageMaker HyperPod**: Purpose-built infrastructure for distributed training of foundation models with up to 40% faster training
- **SageMaker Feature Store**: Centralized repository for storing, sharing, and managing ML features across teams
- **SageMaker Pipelines**: CI/CD service for creating, managing, and executing end-to-end ML workflows
- **SageMaker Clarify**: Bias detection and model explainability service for responsible AI development

- **SageMaker Edge**: Deploy and manage ML models on edge devices with optimization capabilities

Value Proposition

SageMaker AI provides the most comprehensive and fully managed ML platform with choice of tools from low-code to expert-level development environments. It offers 54% lower total cost of ownership compared to self-managed solutions, with built-in governance and security features that meet enterprise requirements:

- The platform uniquely combines traditional ML capabilities with cutting-edge generative AI development, providing access to foundation models while enabling custom model development
- Its fully managed infrastructure eliminates operational overhead, while offering flexible pricing including Savings Plans with up to 64% cost reduction for consistent workloads

Service Integration

SageMaker AI deeply integrates with core AWS services to provide a seamless data-to-model pipeline. It connects with Amazon S3 for scalable data storage, Amazon EMR and AWS Glue for data processing and ETL operations, Amazon Athena for SQL analytics, and Amazon Redshift for data warehousing:

- The platform leverages AWS IAM for security and access control, Amazon CloudWatch for monitoring and logging, and AWS Lambda for serverless integrations
- It also integrates with Amazon Bedrock for foundation model access, Amazon EC2 for compute resources, and supports container deployment through Amazon ECS and EKS

Onboarding and Offboarding

Customers can get started with SageMaker AI through quick setup for individual users with default settings or custom setup for enterprise ML administrators managing organizational deployments. The platform offers multiple entry points including SageMaker Studio for comprehensive ML development, SageMaker Canvas for no-code model building, and SageMaker Studio Lab for learning without requiring an AWS account. New users can access free tier benefits including 250 hours of notebook usage, 50 hours of training, and 125 hours of hosting monthly for the first two months.

Getting Started Guide: https://aws.amazon.com/getting-started/hands-on/build-train-deploy-machine-learning-model-sagemaker/

Data Removal

Data removal follows standard AWS data lifecycle management practices where customers maintain full control over their data. Training data, model artifacts, and notebook files can be deleted through the SageMaker console or APIs. Customers should delete S3 buckets containing model data, terminate SageMaker endpoints, and remove associated resources. IAM roles and policies should be cleaned up, and any shared resources in Feature Store should be properly managed during offboarding.

Backup/Restore and Disaster Recovery

SageMaker provides backup capabilities through integration with AWS native services including S3 versioning for model artifacts and training data. Cross-region disaster recovery can be implemented using Amazon EFS replication for SageMaker domains. Event-driven backup solutions using CloudTrail, EventBridge, and Lambda enable automated metadata capture and restoration workflows for comprehensive disaster recovery planning.

Backup Capabilities

SageMaker supports automated backup through S3 integration for model artifacts and training data with versioning capabilities. Model Registry provides versioning and metadata management for deployed models. The service integrates with AWS Backup for centralized backup management of supported resources, though direct SageMaker-specific backup features focus primarily on data and model artifact preservation.

Disaster Recovery

SageMaker supports disaster recovery through multi-region deployments and cross-region replication capabilities. Custom Amazon EFS instances enable cross-region disaster recovery for SageMaker domains. The service leverages AWS's global infrastructure for regional failover, and automated recovery workflows can be implemented using event-driven architecture patterns with AWS native services.

Recovery Objectives

SageMaker helps meet RPO and RTO objectives through automated model deployment pipelines, multi-region inference endpoints, and real-time model monitoring. The platform supports rapid model redeployment and scaling to handle failover scenarios. Cross-region backup strategies and automated recovery workflows enable customers to achieve aggressive recovery objectives based on their specific business requirements and criticality levels.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/sagemaker/latest/dg/training-plan-quotas.html

FAQ: https://aws.amazon.com/sagemaker/ai/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/sagemaker/latest/dg/index.html

User Guide: https://docs.aws.amazon.com/sagemaker/latest/dg/gs.html

API Reference: https://docs.aws.amazon.com/sagemaker/latest/dg/reference.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/sagemaker-ai/pricing/

Cost Optimization

SageMaker offers unique cost optimization through ML Savings Plans providing up to 64% savings for consistent usage, managed Spot training reducing costs by up to 90%, and automatic model optimization features. The platform provides intelligent instance selection recommendations, automated resource scaling, and usage-based pricing for serverless inference. Customers can leverage inference optimization tools, multi-model endpoints for better resource utilization, and flexible deployment options including asynchronous and batch inference for cost-effective processing.

Savings Considerations

Primary cost savings come from eliminating infrastructure management overhead, using Spot instances for non-time-sensitive training, and implementing ML Savings Plans for predictable workloads. The platform's automatic scaling prevents over-provisioning, while serverless inference options reduce costs for variable traffic patterns:

- Multi-model endpoints enable resource sharing, and built-in optimization tools help rightsize instances
- The fully managed nature eliminates operational costs associated with cluster management, security patching, and infrastructure maintenance typically required in self-managed solutions

Service Name

Amazon Simple Email Service (SES)

Service Overview

Amazon Simple Email Service (SES) is a cost-effective, flexible, and scalable email service that provides an easy way to send and receive email using your own email addresses and domains. It eliminates the complexity and costs of building large-scale email infrastructure by handling server management, network configuration, and IP address reputation. SES supports sending marketing emails, transactional emails, and newsletters, while also enabling email receiving capabilities for developing autoresponders, unsubscribe systems, and customer support ticket generation from incoming emails.

Key Use Cases

- Marketing emails: Send promotional campaigns, special offers, and newsletters to customers
- Transactional emails: Send order confirmations, password resets, and account notifications
- Email receiving and processing: Develop email autoresponders, unsubscribe systems, and customer support ticket generation

Features

- **Email Sending**: Send emails via SMTP interface, API calls, or AWS SDKs with support for formatted and raw email formats
- **Email Receiving**: Receive and process incoming emails with receipt rules, IP filtering, and integration with other AWS services
- **Virtual Deliverability Manager**: Suite of deliverability tools to track and improve email delivery rates and sender reputation
- **Dedicated IP Addresses**: Option for managed or standard dedicated IPs to control sender reputation
- **Mail Manager**: Advanced email traffic management with filtering policies, archiving, and security add-ons
- **Template Support**: Create and manage email templates for personalized bulk messaging
- **Event Publishing**: Track email events like bounces, complaints, and deliveries through CloudWatch and SNS
- **DKIM and Authentication**: Easy DKIM setup and email authentication to improve deliverability

Value Proposition

Amazon SES eliminates infrastructure challenges such as email server management, network configuration, and IP address reputation management that businesses typically face when building large-scale email solutions. It provides cost-effective pay-as-you-go pricing with no upfront costs or long-term contracts, leveraging Amazon's sophisticated email infrastructure built for its own large-scale operations. SES offers seamless integration with other AWS services, high deliverability rates through managed IP pools and authentication features, and scalable infrastructure that can handle varying email volumes from small businesses to enterprise-level operations.

Service Integration

Amazon SES integrates deeply with numerous AWS services including Amazon EC2 for application hosting, AWS Lambda for email processing automation, Amazon S3 for email storage, Amazon SNS for bounce and complaint notifications, Amazon CloudWatch for metrics and monitoring, AWS Identity and Access Management (IAM) for access control, AWS KMS for email encryption, AWS CloudTrail for API logging, Route 53 for DNS management, and Amazon Elastic Beanstalk for email-enabled applications. These integrations enable comprehensive email workflows and automation scenarios.

Onboarding and Offboarding

Getting started with Amazon SES requires signing up for an AWS account and verifying email addresses or domains to prove ownership. New accounts begin in the sandbox environment with limited sending capabilities (200 emails per day, 1 per second) restricted to verified addresses for testing:

- To move to production, customers submit a sending limit increase request through AWS Support Center, providing details about their use case
- The setup process includes configuring sender identities, setting up authentication (DKIM), and optionally configuring dedicated IPs, event publishing, and integration with other AWS services

Getting Started Guide: https://docs.aws.amazon.com/ses/latest/dg/getting-started.html

Data Removal

Amazon SES does not store customer email content by default, as emails are processed and delivered in real-time. For email receiving features, customers can control data retention through their configured storage destinations (S3 buckets). To offboard, customers should delete configuration sets, verified identities, email templates, receipt rules, and any stored email content in connected services like S3. Account closure automatically removes SES configurations and stops billing.

Backup/Restore and Disaster Recovery

Amazon SES is primarily a transactional service that processes emails in real-time without built-in backup capabilities for email content. The service configuration (identities, templates, rules) can be recreated through Infrastructure as Code. For email receiving scenarios, backup depends on the chosen storage destination (typically S3) which provides its own durability and versioning capabilities. SES configurations should be documented and version-controlled for disaster recovery scenarios.

Backup Capabilities

Amazon SES does not provide traditional backup capabilities as it is designed as a real-time email processing service. Email sending is transactional and emails are not stored after processing:

- For email receiving, backup depends on the configured destinations such as S3 buckets which have their own backup and versioning features
- SES works with AWS Backup indirectly through the storage services it integrates with, but not directly as a supported resource

Disaster Recovery

Amazon SES supports disaster recovery through its multi-region availability and the ability to replicate configurations across regions. The service does not directly integrate with AWS Elastic Disaster Recovery as it processes emails in real-time without persistent storage. Disaster recovery planning should include recreating SES configurations, verified identities, and email templates in alternative regions using Infrastructure as Code practices.

Recovery Objectives

Amazon SES helps customers meet RPO and RTO objectives through its multi-region deployment options and real-time processing capabilities. Since emails are processed immediately without storage, RPO is essentially zero for the email service itself. RTO can be minimized by maintaining SES configurations in multiple regions and using automated deployment scripts. For email receiving scenarios, RPO and RTO depend on the configured storage destinations and their respective backup strategies.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/ses/latest/dg/quotas.html

FAQ: https://aws.amazon.com/ses/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/ses/latest/dg/Welcome.html

User Guide: https://docs.aws.amazon.com/ses/latest/DeveloperGuide/Welcome.html

API Reference: https://docs.aws.amazon.com/ses/latest/APIReference/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/ses/pricing/

Cost Optimization

Amazon SES offers several unique cost optimization opportunities including pay-per-use pricing with no minimum fees or contracts, AWS Free Tier providing 3,000 message charges free monthly for the first 12 months, and tiered pricing for Virtual Deliverability Manager that reduces costs for high-volume senders. Additional optimizations include choosing appropriate regions to minimize data transfer costs, using shared IP pools instead of dedicated IPs for lower volumes, and leveraging AWS service integrations to reduce third-party email service costs.

Savings Considerations

Main cost savings considerations include eliminating infrastructure costs for email servers and maintenance, reducing third-party email service fees through competitive per-message pricing ($0.10 per 1,000 emails), and avoiding upfront capital expenditures. Organizations can optimize costs by right-sizing their email volume plans, using the free tier effectively during testing phases, selecting cost-effective storage options for email receiving scenarios, and implementing efficient email practices to minimize bounce rates and improve deliverability, which reduces wasted sending costs.

Service Name

Amazon Simple Notification Service

Service Overview

Amazon Simple Notification Service (Amazon SNS) is a fully managed messaging service that provides message delivery from publishers to subscribers through asynchronous communication. Publishers send messages to a topic, which acts as a logical access point and communication channel, ensuring messages are delivered to multiple subscribers across various platforms. SNS supports both Application-to-Application (A2A) and Application-to-Person (A2P) messaging, enabling flexible communication between applications and directly to mobile phones, email addresses, and other endpoints through various protocols including Amazon SQS, Lambda, HTTP(S), email, SMS, and mobile push notifications.

Key Use Cases

- Application integration through fanout messaging: Messages published to an SNS topic are replicated and pushed to multiple endpoints for parallel asynchronous processing, enabling decoupled microservices architectures
- Application alerts and monitoring: Notifications triggered by predefined thresholds that can be sent to users via SMS, email, and other channels for system monitoring and alerting
- Mobile push notifications: Direct message delivery to mobile applications across various platforms including iOS, Android, and other mobile devices for user engagement
- User notifications: Messages sent to individuals or groups such as e-commerce order confirmations, account updates, and personalized communications

Features

- **Standard and FIFO Topics**: Standard topics support high throughput with best-effort ordering and deduplication, while FIFO topics ensure strict ordering, strict deduplication, and exactly-once processing
- **Message Filtering**: Enables subscribers to receive only relevant notifications by assigning filter policies to topic subscriptions based on message attributes or body content
- **Message Archiving and Replay**: FIFO topics support message archiving for up to 365 days with replay capabilities for message recovery, state replication, and error correction

- **Multi-Protocol Support**: Delivers messages to various endpoints including Amazon SQS, Lambda, HTTP(S), email, SMS, mobile push notifications, and Amazon Data Firehose
- **Server-Side Encryption**: Protects messages using AWS Key Management Service (KMS) with support for both AWS managed and customer managed keys
- **Message Durability**: Ensures message persistence through storage across multiple servers and data centers with delivery retry policies and dead-letter queues

Value Proposition

Amazon SNS offers a fully managed, highly scalable pub/sub messaging platform that eliminates the operational overhead of running messaging infrastructure. Key advantages include: immediate message delivery through push-based communication model, automatic scaling to handle millions of messages per second, extensive integration ecosystem with AWS services and external systems, pay-as-you-go pricing with no upfront costs or minimum commitments, built-in durability and reliability through multi-AZ message storage, advanced features like message filtering and FIFO ordering, and comprehensive security through encryption and VPC endpoints. SNS provides the flexibility to broadcast messages to multiple subscribers simultaneously while maintaining loose coupling between system components.

Service Integration

Amazon SNS integrates deeply with numerous AWS services as both event source and destination. Core integrations include Amazon SQS for reliable message queuing and fanout patterns, AWS Lambda for serverless event processing, Amazon EventBridge for event-driven architectures, AWS Step Functions for workflow orchestration, Amazon Data Firehose for real-time data streaming to analytics services, Amazon Connect for contact center notifications, AWS End User Messaging SMS for two-way SMS capabilities, Amazon SES for email bounce and complaint notifications, and Amazon CloudWatch for monitoring and alerting. SNS also serves as a notification mechanism for various AWS services including EC2 Auto Scaling, CloudFormation, and AWS Backup, creating a comprehensive event-driven ecosystem.

Onboarding and Offboarding

Getting started with Amazon SNS involves creating an AWS account and setting up IAM permissions for administrative access. Users can access SNS through the AWS Management Console, AWS CLI, or AWS SDKs:

- The basic workflow includes: creating an SNS topic as a communication channel, creating subscriptions to define message delivery endpoints, configuring any necessary message filtering or encryption settings, and publishing messages to the topic

- The console provides a graphical interface for basic operations, while CLI and SDKs offer advanced configuration and automation capabilities
- AWS provides comprehensive code examples, tutorials, and getting started guides across multiple programming languages to facilitate rapid onboarding

Getting Started Guide: https://docs.aws.amazon.com/sns/latest/dg/sns-setting-up.html

Data Removal

SNS data removal is straightforward through topic and subscription deletion. When deleting an SNS topic, all associated subscriptions are automatically deleted asynchronously, though unconfirmed subscriptions are automatically removed after 48 hours. Deleted topics cannot be recovered, and any attempts to publish or subscribe to deleted topics result in error messages. Users can delete resources through the AWS Management Console, AWS CLI, or SDKs. For FIFO topics with message archiving enabled, archived messages are also removed when the topic is deleted.

Backup/Restore and Disaster Recovery

Amazon SNS provides built-in resilience through multi-AZ message storage and automatic replication across multiple servers and data centers within a region. FIFO topics offer message archiving capabilities for up to 365 days, enabling message recovery and replay functionality. SNS topics and configurations can be recreated in other regions using Infrastructure as Code templates. The service itself doesn't require traditional backup as it's fully managed, but topic configurations, subscription settings, and access policies should be documented or stored in version control systems.

Backup Capabilities

SNS doesn't integrate directly with AWS Backup as it's a messaging service rather than a data storage service. However, FIFO topics provide built-in message archiving that stores messages for up to 365 days with replay capabilities:

- The service ensures message durability through automatic replication across multiple servers and availability zones
- Topic configurations and policies can be backed up through Infrastructure as Code practices using CloudFormation or similar tools

Disaster Recovery

SNS doesn't directly integrate with AWS Elastic Disaster Recovery as it's a messaging service, but it supports multi-region disaster recovery strategies. Organizations can create identical SNS topics and subscriptions in multiple regions and implement cross-region failover logic in their applications. The service's built-in durability features, combined with multi-region deployment patterns and AWS Application Recovery Controller for traffic

routing, enable comprehensive disaster recovery solutions with configurable RTO and RPO objectives.

Recovery Objectives

Amazon SNS helps meet aggressive RTO and RPO objectives through its push-based, real-time message delivery architecture and built-in durability features. FIFO topics with message archiving can achieve near-zero RPO for critical messaging workflows through message replay capabilities. Multi-region deployments can provide sub-minute RTO through automated failover mechanisms. The service's automatic scaling and multi-AZ redundancy ensure high availability during regional failures.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/general/latest/gr/sns.html

FAQ: https://aws.amazon.com/sns/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/sns/latest/dg/welcome.html

User Guide: https://docs.aws.amazon.com/sns/latest/dg/welcome.html

API Reference: https://docs.aws.amazon.com/sns/latest/api/

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/sns/pricing/

Cost Optimization

SNS offers several cost optimization strategies: utilize message batching to reduce API call costs by publishing up to 10 messages in a single request, implement attribute-based message filtering to reduce unnecessary deliveries and associated charges, leverage the free tier offering 1 million publishes and notifications monthly for new accounts, optimize message size as each 64KB chunk is billed separately, use appropriate topic types with FIFO topics being more expensive than standard topics, implement efficient retry policies to avoid excessive delivery attempts, and consider regional deployment strategies to minimize data transfer costs. Message filtering at the subscription level prevents unnecessary processing and delivery charges.

Savings Considerations

Primary cost savings opportunities include: taking advantage of the generous free tier limits for development and low-volume production workloads, implementing message filtering to eliminate unwanted deliveries and reduce notification costs, using standard topics instead of FIFO when strict ordering isn't required, optimizing message payload sizes to stay under 64KB chunks, leveraging batching capabilities to reduce API request charges, monitoring and optimizing delivery retry policies to prevent excessive failed delivery attempts, using appropriate subscription protocols with some being more cost-effective than others, and implementing proper message deduplication to avoid duplicate processing costs. Cross-region data transfer should be minimized through strategic regional deployments.

Service Name

Amazon Simple Queue Service (SQS)

Service Overview

Amazon Simple Queue Service (SQS) is a fully managed message queuing service that enables decoupling and integration of distributed software systems and components. SQS provides secure, durable, and highly available hosted queues for transmitting messages between applications. The service offers two queue types: Standard queues providing unlimited throughput and best-effort ordering with at-least-once delivery, and FIFO queues delivering exactly-once processing with first-in-first-out ordering. SQS eliminates the complexity and overhead of managing message brokers while ensuring reliable message delivery across distributed architectures. The service supports server-side encryption, dead-letter queues, and seamless integration with other AWS services, making it ideal for microservices architectures and event-driven applications.

Key Use Cases

- Decoupling microservices: Enable loose coupling between application components for improved fault tolerance and scalability
- Work queues: Distribute tasks across multiple worker processes for parallel processing and load balancing
- Buffer and batch operations: Handle traffic spikes by queuing requests for batch processing during off-peak times
- Request offloading: Move time-consuming operations to background processing to improve application responsiveness
- Event-driven architectures: Enable asynchronous communication between services in serverless and container-based applications
- Order management: Process e-commerce transactions and financial operations requiring strict message ordering with FIFO queues

Features

- **Standard and FIFO Queues**: Standard queues offer unlimited throughput with at-least-once delivery, while FIFO queues provide exactly-once processing with guaranteed ordering
- **Server-Side Encryption**: Encrypt messages at rest using SQS-managed keys or AWS KMS customer-managed keys for enhanced security
- **Dead-Letter Queues**: Handle failed messages by automatically moving them to a separate queue after maximum receive attempts
- **Long Polling**: Reduce costs and improve performance by waiting for messages to arrive instead of frequent empty requests

- **Message Batching**: Send, receive, and delete up to 10 messages in a single API call to improve throughput and reduce costs
- **Visibility Timeout**: Prevent multiple consumers from processing the same message by making it temporarily invisible during processing
- **Message Retention**: Store messages for up to 14 days with configurable retention periods from 1 minute to 14 days
- **Fair Queues**: Limit per-tenant consumption rates to prevent noisy neighbor scenarios in multi-tenant applications

Value Proposition

Amazon SQS provides significant advantages over self-managed message brokers and alternative solutions. The fully managed nature eliminates infrastructure maintenance, patching, and scaling concerns while providing 99.9% availability SLA:

- Cost efficiency comes from pay-per-use pricing with no minimum fees and a generous free tier of 1 million requests monthly
- SQS offers superior integration with the AWS ecosystem, enabling seamless workflows with Lambda, SNS, S3, and other services
- The service automatically scales to handle any load without provisioning, supports both high-throughput standard queues and ordered FIFO processing, and provides enterprise-grade security with encryption at rest and in transit
- Unlike traditional message brokers, SQS requires no capacity planning or cluster management, making it ideal for cloud-native applications seeking reliability and operational simplicity

Service Integration

Amazon SQS deeply integrates with core AWS services to enable comprehensive cloud architectures. AWS Lambda functions can be triggered directly by SQS messages for serverless event processing:

- Amazon SNS provides fan-out capabilities by delivering messages to multiple SQS queues simultaneously
- Amazon S3 integration enables handling large messages by storing message content in S3 and passing references through SQS
- Amazon DynamoDB works with SQS for reliable data processing workflows
- Amazon EC2 and ECS applications can consume messages for distributed processing
- AWS CloudWatch provides monitoring and alerting for queue metrics
- Amazon EventBridge enables complex event routing patterns
- AWS IAM controls access permissions, while AWS KMS manages encryption keys
- Amazon VPC endpoints allow private connectivity without internet gateways

- AWS CloudFormation and CDK support infrastructure as code deployments

Onboarding and Offboarding

Getting started with Amazon SQS requires an AWS account and appropriate IAM permissions. Users begin by creating their first queue through the AWS Management Console, AWS CLI, or SDKs, choosing between Standard or FIFO queue types based on their requirements:

- The console provides an intuitive interface for queue creation, configuration, and message management
- After creating a queue, customers can immediately start sending and receiving messages using the provided queue URL
- AWS provides comprehensive SDKs for popular programming languages including Java, Python, JavaScript, .NET, Go, and others
- The service includes extensive documentation, tutorials, and code examples to accelerate development
- Best practices include configuring dead-letter queues, setting appropriate visibility timeouts, enabling server-side encryption, and implementing proper IAM policies
- AWS offers free tier usage and transparent pricing to minimize initial costs during evaluation and development phases

Getting Started Guide:
https://docs.aws.amazon.com/AWSSimpleQueueService/latest/SQSDeveloperGuide/sqs-getting-started.html

Data Removal

Offboarding from Amazon SQS involves deleting queues and their contained messages through the AWS Management Console, AWS CLI, or APIs. When a queue is deleted, all messages within it are permanently removed and cannot be recovered. Customers should ensure all important messages are processed or backed up before deletion. Queue deletion is immediate and irreversible. For applications using SQS, customers must update their code to remove queue references and stop message processing. IAM policies and roles associated with SQS should be cleaned up to maintain security hygiene.

Backup/Restore and Disaster Recovery

Amazon SQS provides built-in disaster recovery through its distributed architecture across multiple Availability Zones within a region. Messages are automatically stored redundantly across multiple servers for durability. However, SQS does not provide traditional backup capabilities for message content. For disaster recovery across regions, customers must implement their own cross-region message replication using application logic, Lambda functions, or other AWS services. The service itself is designed for high availability with

automatic failover capabilities, eliminating the need for traditional backup and restore mechanisms for the queueing infrastructure.

Backup Capabilities

Amazon SQS does not integrate with AWS Backup service as it is designed as a transient messaging service rather than a data storage service. Messages are intended to be processed and deleted, not permanently stored:

- The service provides message retention capabilities for up to 14 days, but this is for operational purposes rather than backup
- For long-term message archival, customers should implement their own solutions using Amazon S3, DynamoDB, or other storage services in conjunction with SQS processing workflows

Disaster Recovery

Amazon SQS does not directly integrate with AWS Elastic Disaster Recovery service since it is a managed messaging service with built-in high availability. SQS automatically provides redundancy across multiple Availability Zones within a region:

- For cross-region disaster recovery scenarios, customers must architect their own solutions using multiple SQS queues across regions, combined with application-level logic or Lambda functions to replicate messages
- The service's distributed nature and regional availability provide inherent disaster recovery capabilities for single AZ failures

Recovery Objectives

Amazon SQS helps customers achieve aggressive RPO and RTO objectives through its inherent design. The service provides near-zero RPO for regional disasters due to synchronous replication across multiple Availability Zones. RTO is typically measured in seconds for AZ-level failures due to automatic failover capabilities. For message processing workflows, customers can achieve sub-minute RTOs by implementing multi-AZ consumer deployment patterns. Cross-region scenarios require customer-implemented replication strategies but can still achieve RPO and RTO objectives in the minutes range depending on application architecture and automation sophistication.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/AWSSimpleQueueService/latest/SQSDeveloperGuide/sqs-quotas.html

FAQ: https://aws.amazon.com/sqs/faqs/

Technical Requirements

Service Documentation:
https://docs.aws.amazon.com/AWSSimpleQueueService/latest/SQSDeveloperGuide/welcome.html

User Guide:
https://docs.aws.amazon.com/AWSSimpleQueueService/latest/SQSDeveloperGuide/sqs-getting-started.html

API Reference:
https://docs.aws.amazon.com/AWSSimpleQueueService/latest/APIReference/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/sqs/pricing/

Cost Optimization

Amazon SQS offers several unique cost optimization opportunities including message batching to reduce API request costs by up to 90%, long polling to eliminate empty receive requests and associated charges, and efficient queue management to avoid unnecessary storage costs. The service's pay-per-use model with no minimum fees makes it inherently cost-effective:

- Customers can optimize by using standard queues for high-throughput scenarios where ordering is not critical, implementing proper message lifecycle management to avoid retention charges, and leveraging the free tier of 1 million requests monthly
- FIFO queues should only be used when strict ordering is required due to higher per-request costs
- Proper dead-letter queue configuration prevents infinite retry loops that generate unnecessary charges

Savings Considerations

Primary cost savings with Amazon SQS come from eliminating message broker infrastructure, licensing, and operational overhead. The fully managed nature removes costs associated with server provisioning, patching, monitoring, and scaling:

- Batch operations can reduce request costs by processing up to 10 messages per API call
- Long polling reduces costs by eliminating empty receive requests

- The service's automatic scaling prevents over-provisioning costs common with traditional message brokers
- Pay-per-use pricing ensures customers only pay for actual usage without capacity commitments
- Integration with other AWS services reduces data transfer costs within the same region
- The generous free tier provides substantial cost savings for development, testing, and low-volume production workloads
- Proper queue lifecycle management and monitoring help identify and eliminate unused resources

Service Name

Amazon Simple Storage Service (S3)

Service Overview

Amazon S3 is a scalable object storage service that offers industry-leading scalability, data availability, security, and performance. S3 stores data as objects within buckets and provides 99.999999999% (11 9's) durability. It offers strong read-after-write consistency for PUT and DELETE requests across all AWS Regions. The service supports unlimited storage capacity with individual objects up to 5 TB in size, making it suitable for storing and retrieving any amount of data from anywhere on the web through a simple web services interface.

Key Use Cases

- Data lakes and big data analytics: Central repository for structured and unstructured data with integration to analytics services
- Backup and restore: Reliable, scalable backup storage with cross-region replication capabilities
- Website and mobile application hosting: Static website hosting and content distribution for web and mobile apps
- Archive and long-term storage: Cost-effective archival using Glacier storage classes for compliance and regulatory requirements
- Enterprise application storage: Scalable storage backend for enterprise applications and microservices
- IoT device data storage: Collect and store data from millions of IoT devices with event-driven processing
- Media and content distribution: Store and deliver video, images, and other media content globally

Features

- **Multiple Storage Classes**: Eight storage classes including S3 Standard, Express One Zone, Intelligent-Tiering, Standard-IA, One Zone-IA, and Glacier classes for different access patterns and cost optimization
- **S3 Lifecycle**: Automate object transitions between storage classes and deletion based on defined rules to optimize costs
- **S3 Replication**: Cross-Region Replication (CRR) and Same-Region Replication (SRR) for data redundancy, compliance, and disaster recovery
- **S3 Versioning**: Preserve, retrieve, and restore every version of every object for data protection against accidental deletion or modification

- **S3 Object Lock**: Write-once-read-many (WORM) model to prevent objects from being deleted or overwritten for compliance requirements
- **Access Management**: Comprehensive access control through IAM, bucket policies, Access Control Lists (ACLs), and S3 Block Public Access
- **Encryption**: Server-side encryption (SSE-S3, SSE-KMS, SSE-C) and client-side encryption options for data protection
- **S3 Batch Operations**: Perform large-scale batch operations on billions of objects with a single API request
- **Event Notifications**: Trigger workflows and processing through integration with Lambda, SQS, and SNS when objects are created, deleted, or modified
- **Analytics and Insights**: S3 Storage Lens, Storage Class Analysis, and S3 Inventory for visibility into storage usage and cost optimization

Value Proposition

Amazon S3 offers unmatched scalability with virtually unlimited storage capacity and the ability to scale from gigabytes to exabytes seamlessly. It provides industry-leading 99.999999999% durability and 99.99% availability, ensuring data protection across multiple Availability Zones:

- S3's comprehensive security features include encryption, access controls, and compliance certifications like PCI DSS
- The service offers significant cost optimization through multiple storage classes, intelligent tiering, and lifecycle policies
- S3 integrates natively with over 100 AWS services and provides global accessibility through REST APIs, SDKs, and management console
- The pay-as-you-go pricing model with no upfront costs or minimum commitments makes it cost-effective for organizations of all sizes

Service Integration

Amazon S3 integrates deeply with core AWS services to enable comprehensive cloud solutions. Compute services like EC2, Lambda, and EMR use S3 for data storage and processing:

- Analytics services including Athena, Redshift Spectrum, and QuickSight query data directly from S3
- Database services like RDS and DynamoDB back up to S3
- Content delivery through CloudFront uses S3 as origin storage
- Security services like CloudTrail store logs in S3, while IAM manages S3 access permissions
- Data transfer services including Snow Family, DataSync, and Transfer Family move data to/from S3

- Machine learning services like SageMaker and Rekognition process data stored in S3
- AWS Backup provides centralized backup management for S3 and other services

Onboarding and Offboarding

Getting started with Amazon S3 requires signing up for an AWS account and creating an IAM user with appropriate permissions. Customers can immediately create their first S3 bucket through the AWS Management Console, CLI, or SDKs:

- The process involves selecting a unique bucket name, choosing a region, and configuring access settings
- Objects can be uploaded using the console interface, programmatic APIs, or bulk transfer tools
- S3 provides extensive documentation, tutorials, and code examples for all major programming languages
- AWS offers free tier access with 5 GB of S3 Standard storage, 20,000 GET requests, and 2,000 PUT requests per month for 12 months, enabling customers to explore S3 capabilities without initial costs

Getting Started Guide:
https://docs.aws.amazon.com/AmazonS3/latest/userguide/GetStartedWithS3.html

Data Removal

S3 offboarding involves systematic data deletion and account cleanup. Customers must delete all objects within buckets before deleting buckets themselves. Versioned objects require deletion of all versions and delete markers. Multi-part uploads should be aborted and lifecycle policies can automate cleanup. Cross-region replicated data must be removed from all regions. Access policies, bucket policies, and IAM permissions should be revoked. Customers can use S3 Inventory to identify all objects and S3 Batch Operations for bulk deletion operations.

Backup/Restore and Disaster Recovery

S3 provides comprehensive backup and disaster recovery capabilities through Cross-Region Replication (CRR), versioning, and integration with AWS Backup service. CRR automatically replicates objects to different AWS regions for geographic redundancy. S3 versioning protects against accidental deletion or corruption by maintaining multiple object versions. AWS Backup offers centralized, policy-driven backup management across AWS services with automated scheduling and cross-region backup copying for enhanced disaster recovery protection.

Backup Capabilities

S3 offers native backup capabilities including object versioning, Cross-Region Replication, and lifecycle policies for automated backup management. The service integrates fully with AWS Backup for centralized backup policies and management:

- S3's 99.999999999% durability provides inherent data protection by storing multiple copies across different facilities
- Point-in-time recovery is available through versioning, and automated backup scheduling can be configured through lifecycle rules and AWS Backup policies

Disaster Recovery

S3 supports comprehensive disaster recovery through Cross-Region Replication for geographic redundancy and AWS Elastic Disaster Recovery for application-level protection. Multi-Region Access Points provide automatic failover capabilities:

- S3's global infrastructure across multiple AWS regions enables robust disaster recovery architectures
- While S3 doesn't directly integrate with AWS Elastic Disaster Recovery as a replication target, it serves as a critical storage foundation for disaster recovery solutions and backup storage for applications protected by disaster recovery services

Recovery Objectives

S3 helps customers achieve aggressive RPO and RTO objectives through multiple mechanisms. Cross-Region Replication can achieve RPOs of 15 minutes or less for most objects. S3 Replication Time Control provides 99.99% of objects replicated within 15 minutes with SLA backing. Versioning enables near-zero RPO for data corruption scenarios. Multi-Region Access Points enable automatic failover reducing RTO to minutes. The combination of these features allows customers to design solutions meeting even sub-minute RTO requirements for critical applications.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/general/latest/gr/s3.html

FAQ: https://aws.amazon.com/s3/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/AmazonS3/latest/userguide/Welcome.html

User Guide: https://docs.aws.amazon.com/AmazonS3/latest/userguide/

API Reference: https://docs.aws.amazon.com/AmazonS3/latest/API/

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/s3/pricing/

Cost Optimization

S3 provides exceptional cost optimization through S3 Intelligent-Tiering, which automatically moves data between access tiers based on usage patterns, delivering up to 68% cost savings without performance impact or operational overhead. Eight distinct storage classes serve different access patterns from high-performance S3 Express One Zone to ultra-low-cost S3 Glacier Deep Archive:

- Lifecycle policies automate transitions to lower-cost storage classes
- S3 Storage Class Analysis identifies optimization opportunities, while S3 Storage Lens provides organization-wide cost visibility
- Request charges can be optimized through bulk operations and appropriate API selection
- Data transfer costs can be reduced through CloudFront caching and careful region selection

Savings Considerations

Primary cost savings come from choosing appropriate storage classes for access patterns - Glacier classes can provide up to 95% cost savings for archival data. S3 Intelligent-Tiering eliminates the complexity of lifecycle management while delivering automatic cost optimization:

- Lifecycle policies can automatically delete incomplete multipart uploads and expired object versions
- Regional data storage location affects costs, with some regions offering lower pricing
- Bulk operations through S3 Batch Operations reduce per-request charges
- Using S3 Transfer Acceleration judiciously and optimizing data transfer patterns can significantly reduce egress costs
- Reserved capacity pricing for predictable workloads and savings plans can provide additional discounts for committed usage

Service Name

Amazon Simple Workflow Service (SWF)

Service Overview

Amazon Simple Workflow Service (SWF) is a fully managed workflow service that enables developers to build, run, and scale background jobs with parallel or sequential steps. It coordinates work across distributed components and manages task state, handling inter-task dependencies, scheduling, and concurrency in application workflows. SWF stores tasks, assigns them to workers when ready, tracks progress, and maintains state including task completion details. Workers can run on cloud infrastructure or on-premises, performing long-running tasks that may fail, timeout, or require restarts with varying throughput and latency.

Key Use Cases

- Video encoding: Processing media files through multiple encoding stages with parallel or sequential workflows
- Data center migration: Orchestrating complex migration processes with multiple dependent steps and rollback capabilities
- E-commerce order processing: Coordinating order verification, payment processing, inventory management, and shipping workflows
- Financial transactions: Managing multi-step financial processes with strict consistency and audit requirements
- Product catalogs with human workers: Workflows requiring human intervention for approvals, reviews, or manual processing steps

Features

- **Task Coordination**: Manages inter-task dependencies, scheduling, and concurrency across distributed components
- **State Management**: Maintains durable execution state allowing for resilient applications and independent component failures
- **Worker Management**: Assigns tasks to workers running on cloud or on-premises infrastructure with polling-based communication
- **Decision Tasks**: Enables workflow logic through deciders that analyze execution history and make routing decisions
- **Activity Tasks**: Supports activity workers performing business logic with automatic retry and failure handling
- **Lambda Tasks**: Direct integration with AWS Lambda functions for serverless workflow components

- **Child Workflows**: Supports nested workflows for complex process decomposition and reusability
- **Workflow History**: Maintains complete execution history for audit, debugging, and state reconstruction

Value Proposition

Amazon SWF provides a fully managed platform for building resilient distributed applications without worrying about underlying infrastructure complexities. It offers durable state management, automatic retry mechanisms, and complete execution history tracking:

- SWF enables independent scaling and deployment of application components while maintaining strong consistency
- The service supports both cloud and on-premises workers, providing flexibility in deployment architectures
- With built-in fault tolerance and no single point of failure, SWF ensures business-critical workflows continue operating even when individual components fail, making it ideal for long-running, mission-critical processes

Service Integration

Amazon SWF integrates with numerous AWS services to provide comprehensive workflow capabilities. It works with AWS Lambda for serverless task execution, Amazon SNS for notifications and human task workflows, Amazon SQS for message queuing, and Amazon DynamoDB for external data storage:

- Integration with Amazon S3 enables large data processing workflows, while IAM provides access control and security
- CloudWatch offers monitoring and metrics, CloudTrail provides audit logging, and the service supports multiple AWS SDKs (Java, .NET, Node.js, PHP, Python, Ruby)
- The AWS Flow Framework for Java provides enhanced programming conveniences for workflow development

Onboarding and Offboarding

Getting started with Amazon SWF involves creating an AWS account and accessing the service through the console, SDKs, or APIs. Customers begin by registering domains, which provide logical groupings for workflows and activities:

- Next, they define and register workflow types and activity types, implement activity workers and deciders, and start workflow executions
- The service offers sample workflows for testing and learning
- AWS provides comprehensive documentation, tutorials, and the AWS Flow Framework for Java to simplify development

- The free tier includes 1,000 workflow executions, 10,000 tasks, and 30,000 workflow-days monthly to support initial exploration and development

Getting Started Guide:
https://docs.aws.amazon.com/amazonswf/latest/developerguide/welcome.html

Data Removal

Amazon SWF automatically manages data retention with completed workflow executions retained for a limited period before automatic deletion. To offboard, customers stop starting new workflow executions, allow existing executions to complete or terminate them manually. Domain deprecation prevents new registrations while preserving existing executions. Data removal occurs through natural retention policies or explicit termination of active executions. No additional data deletion process is required as the service handles cleanup automatically.

Backup/Restore and Disaster Recovery

Amazon SWF provides inherent disaster recovery through its distributed, multi-AZ architecture but does not offer traditional backup capabilities. Workflow execution state is maintained durably by the service itself. For disaster recovery, customers can replicate workflow definitions and activity types across regions. The service's stateless worker model allows for easy failover to different regions. Critical application data should be backed up separately using integrated services like Amazon S3 or DynamoDB.

Backup Capabilities

Amazon SWF does not provide traditional backup capabilities as it maintains workflow state internally through its managed infrastructure. The service automatically handles state persistence and durability:

- Workflow definitions, activity types, and execution history are maintained by the service but cannot be directly backed up by customers
- For application data backup, customers must use external storage services like S3 or DynamoDB and implement appropriate backup strategies for those services

Disaster Recovery

Amazon SWF supports disaster recovery through its distributed architecture across multiple Availability Zones within a region but does not integrate with AWS Elastic Disaster Recovery. For cross-region disaster recovery, customers must manually replicate workflow and activity type registrations across regions:

- The stateless nature of workers enables rapid failover to alternative regions
- Complete disaster recovery requires coordination with other AWS services storing application data and implementing cross-region replication strategies

Recovery Objectives

Amazon SWF helps achieve RPO and RTO objectives through its durable state management and distributed architecture. The service provides near-zero RPO for workflow state within a region through automatic persistence. RTO depends on worker restart times and cross-region failover implementation. Built-in retry mechanisms and fault tolerance reduce recovery time for individual task failures. For optimal recovery objectives, customers should implement workers across multiple regions and maintain workflow type registrations in backup regions.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/amazonswf/latest/developerguide/swf-dg-limits.html

FAQ: https://aws.amazon.com/swf/faqs/

Technical Requirements

Service Documentation:
https://docs.aws.amazon.com/amazonswf/latest/developerguide/index.html

User Guide:
https://docs.aws.amazon.com/amazonswf/latest/developerguide/welcome.html

API Reference:
https://docs.aws.amazon.com/amazonswf/latest/apireference/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/swf/pricing/

Cost Optimization

Amazon SWF offers unique cost optimization through its pay-per-use model based on workflow executions, tasks, and data transfer rather than infrastructure provisioning. The generous free tier (1,000 executions, 10,000 tasks monthly) supports development and small workloads:

- Cost scales directly with usage without minimum commitments
- Efficient workflow design reduces task counts and execution time
- Regional selection impacts data transfer costs
- Long-running workflows are charged per 24-hour period, making efficient execution design crucial

- The stateless worker model allows dynamic scaling based on demand without paying for idle resources

Savings Considerations

Primary cost savings come from efficient workflow design that minimizes task count and execution duration. Batch processing reduces per-task overhead costs:

- Regional consolidation minimizes data transfer charges between services
- The free tier provides substantial savings for development and testing environments
- Compared to building custom workflow infrastructure, SWF eliminates server management, scaling, and high availability costs
- Automatic state management and retry mechanisms reduce operational overhead
- The service's durability eliminates costs associated with building fault-tolerant infrastructure
- For high-volume workloads, the per-task pricing model provides predictable costs without capacity planning complexity

Service Name

Amazon Textract

Service Overview

Amazon Textract is a machine learning service that automatically extracts text, handwriting, layout elements, and structured data from scanned documents and images. It goes beyond simple optical character recognition (OCR) to identify, understand, and extract data from forms and tables without requiring templates or configuration. The service uses deep-learning technology developed by Amazon's computer vision scientists and provides APIs for document analysis, expense processing, ID document analysis, lending document processing, and custom queries. Amazon Textract returns extracted data with bounding box coordinates and confidence scores, making it available to developers without requiring machine learning expertise.

Key Use Cases

- Creating intelligent search indexes: Using Amazon Textract to create libraries of text detected in image and PDF files for searchable document repositories
- Automating data capture from forms: Extracting structured data from forms and building extraction capabilities into existing business workflows for user-submitted data
- Accelerating data capture and normalization: Extracting text and tabular data from financial documents, research reports, medical notes, and other varied document sources
- Intelligent text extraction for NLP: Providing controlled text grouping as input for natural language processing applications with words, lines, and table cell organization
- Automating document classification and processing: Using Analyze Lending API to automatically classify lending documents and route them to appropriate analysis operations

Features

- **Optical Character Recognition (OCR)**: Automatically detects printed and handwritten text from documents and images, handling various fonts, styles, and noisy or distorted text
- **Form Extraction**: Detects key-value pairs in document images automatically while retaining context without manual intervention
- **Table Extraction**: Preserves composition of data stored in tables during extraction, maintaining relationships between rows and columns

- **Query-based Extraction**: Allows specification of data to extract using natural language questions without knowing document structure
- **Custom Queries**: Enables customization of pretrained Queries feature using business-specific documents with as few as ten sample documents
- **Signature Detection**: Provides ability to detect signatures on any document or image with location and confidence scores
- **Layout Analysis**: Extracts layout elements such as paragraphs, titles, lists, headers, footers from documents
- **Analyze Expense**: Extracts vendor names, invoice numbers, item prices, total amounts from invoices and receipts with normalized key names
- **Analyze ID**: Extracts information from US government identity documents like passports and driver's licenses without templates
- **Analyze Lending**: Managed API that automates extraction from mortgage loan documents with prebuilt ML models for classification and splitting

Value Proposition

Amazon Textract offers compelling advantages over traditional OCR solutions and manual data entry processes. It eliminates the complexity of building text detection capabilities by providing powerful and accurate analysis through simple APIs, requiring no computer vision or deep learning expertise:

- The service provides scalable document analysis that can process millions of documents quickly to accelerate decision making
- Unlike basic OCR solutions that extract simple text, Textract maintains relationships between data elements like key-value pairs and table structures
- It offers low-cost, pay-per-use pricing with no minimum fees or upfront commitments, plus a free tier for getting started
- The service is always learning from new data with continual feature additions, and supports both synchronous processing for low-latency applications and asynchronous operations for multi-page documents

Service Integration

Amazon Textract integrates seamlessly with core AWS services to enable comprehensive document processing workflows. It integrates with Amazon S3 for document storage and retrieval, supporting both input and output operations:

- Amazon SNS and Amazon SQS are used for asynchronous processing notifications and job completion management
- AWS Lambda enables serverless processing workflows and automated document processing triggers
- Amazon Cognito provides authentication for web and mobile applications using Textract

- The service works with AWS CloudTrail for API call logging and monitoring
- Amazon EventBridge can trigger document processing workflows
- AWS Step Functions orchestrate complex document processing pipelines
- Amazon DynamoDB and Amazon RDS store extracted data and metadata
- AWS IAM manages access control and permissions
- The service also supports VPC endpoints through AWS PrivateLink for private connectivity without internet gateways

Onboarding and Offboarding

Getting started with Amazon Textract requires an AWS account and basic setup steps. Customers can begin by creating an AWS account and configuring IAM users with appropriate Textract permissions:

- The AWS CLI and SDKs need to be installed and configured with access credentials
- Users can immediately try the service through the Amazon Textract Management Console with their own documents, or explore the demonstration features
- The service offers comprehensive documentation including a developer guide, API reference, and step-by-step getting started guide
- Code samples and tutorials are available on GitHub to help developers integrate Textract into applications
- The service provides SDKs for multiple programming languages including JavaScript, Python, Java, and others
- A free tier allows processing up to 1,000 pages monthly with certain APIs, enabling risk-free evaluation and learning

Getting Started Guide: https://docs.aws.amazon.com/textract/latest/dg/getting-started.html

Data Removal

Amazon Textract provides clear data removal processes for customer offboarding. Document and image inputs are not stored by Amazon Textract unless explicitly requested by customers. When data is stored, customers can delete images and documents using the AWS Management Console or AWS CLI. For Custom Queries adapters, customers can delete both individual adapter versions and entire adapters using the DeleteAdapter and DeleteAdapterVersion API operations, which remove all associated data and ARNs. Customer content used for adapter generation is deleted once training completes and is not retained for debugging purposes. Customers retain full ownership of their content throughout the process.

Backup/Restore and Disaster Recovery

Amazon Textract operates as a managed service where AWS handles infrastructure backup and disaster recovery. The service itself does not require customer-managed backups as it is stateless for processing operations. However, customers should implement backup strategies for their input documents stored in Amazon S3 and any extracted data stored in databases. AWS Backup can be used to protect associated resources like S3 buckets containing documents and RDS databases storing extracted information. The service provides high availability through AWS's global infrastructure with automatic failover capabilities built into the managed service architecture.

Backup Capabilities

Amazon Textract does not have traditional backup capabilities as it is a stateless processing service. The service processes documents and returns results without storing customer data long-term unless explicitly configured:

- For Custom Queries adapters, the trained models are maintained by AWS but can be deleted by customers when no longer needed
- Customers should implement backup strategies for their source documents in S3 and extracted data in downstream storage systems
- AWS Backup can protect related resources but does not directly backup Textract processing results since they are transient

Disaster Recovery

Amazon Textract supports disaster recovery through AWS's global infrastructure and multi-region availability. The service is available in multiple AWS regions allowing customers to process documents in different geographic locations for redundancy:

- As a managed service, AWS handles the underlying infrastructure disaster recovery including automatic failover and data replication
- Customers can implement cross-region strategies by processing documents in multiple regions and storing results in geographically distributed storage systems
- AWS Elastic Disaster Recovery does not directly apply since Textract is a stateless managed service

Recovery Objectives

Amazon Textract helps customers achieve recovery objectives through its managed service architecture and multi-region availability. The service provides near real-time processing for single-page documents, supporting low RTO requirements. For batch processing scenarios, asynchronous operations can be restarted quickly in alternate regions if needed. Since the service is stateless, RPO objectives depend on backing up source documents and extracted results rather than the processing service itself.

Customers can achieve sub-hour RTO by maintaining document repositories in multiple regions and using cross-region replication for critical processing workflows.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/general/latest/gr/textract.html

FAQ: https://aws.amazon.com/textract/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/textract/latest/dg/what-is.html

User Guide: https://docs.aws.amazon.com/textract/latest/dg/what-is.html

API Reference: https://docs.aws.amazon.com/textract/latest/dg/API_Reference.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/textract/pricing/

Cost Optimization

Amazon Textract offers several unique cost optimization opportunities through its pay-per-page pricing model. Customers can optimize costs by preprocessing documents to remove unnecessary pages, using appropriate API features (basic OCR vs:

- advanced analysis), and implementing confidence score thresholds to reduce manual review costs
- Asynchronous processing enables batch operations that can be more cost-effective than real-time processing for large volumes
- The service offers volume-based pricing tiers with reduced costs after one million pages for most APIs
- Custom Queries can improve extraction accuracy, reducing downstream processing costs and manual review requirements
- Customers should evaluate which combination of Features (Forms, Tables, Queries) they actually need rather than using all features by default

Savings Considerations

Main cost savings with Amazon Textract come from eliminating manual data entry and document processing labor costs. The service can process thousands of documents per hour compared to manual processing, providing significant operational savings:

- Volume-based pricing provides cost reductions for high-usage scenarios, with prices decreasing after processing one million pages monthly
- The free tier offers 1,000 pages monthly for text detection, 100 pages for document analysis, and other allowances for evaluation without cost
- Preprocessing optimization like document quality improvement can increase extraction accuracy, reducing costly manual review processes
- Using appropriate APIs for specific use cases (Expense API for invoices vs
- general document analysis) provides better cost efficiency and accuracy than generic approaches

Service Name

Amazon Timestream

Service Overview

Amazon Timestream is a fully managed, serverless time series database service designed for storing and analyzing trillions of time-series data points daily. It offers two engine options: Timestream for LiveAnalytics for real-time analytics with fast SQL queries, and Timestream for InfluxDB for open-source compatibility. The service automatically manages data lifecycle by keeping recent data in memory for fast access and moving historical data to cost-optimized magnetic storage. It provides built-in time series analytics functions, simplified data access across storage tiers, and seamless integration with AWS services for data collection, visualization, and machine learning applications.

Key Use Cases

- IoT sensor data storage and analysis: Collecting and analyzing massive volumes of sensor data from connected devices with real-time monitoring capabilities
- Industrial telemetry management: Managing operational data from manufacturing equipment, power systems, and infrastructure monitoring with predictive maintenance insights
- Application performance monitoring: Tracking application metrics, user interactions, and system performance data for real-time insights and anomaly detection
- Infrastructure monitoring: Collecting and analyzing system metrics, logs, and operational data from servers, networks, and cloud resources for observability

Features

- **Serverless Auto-scaling**: Fully managed service with automatic capacity adjustment based on workload demands
- **Dual Storage Architecture**: Memory store for high-throughput writes and fast queries, magnetic store for cost-effective long-term storage
- **Built-in Time Series Functions**: SQL with time series-specific extensions for analytics, aggregations, and data processing
- **Timestream Compute Units (TCUs)**: Serverless compute capacity with 4 vCPUs and 16 GB memory per unit for predictable query performance
- **Multi-AZ Replication**: Data automatically replicated across three Availability Zones for high availability and durability
- **Data Lifecycle Management**: User-configurable policies for automatic data tiering between memory and magnetic storage

- **Always Encrypted**: Data encrypted at rest and in transit with AWS KMS integration for customer-managed keys
- **Multi-measure Records**: Store multiple measures in single table rows to reduce storage costs and improve query performance

Value Proposition

Amazon Timestream provides significant advantages over traditional relational databases and self-managed solutions for time-series workloads. It delivers up to 10x cost reduction compared to relational databases while handling trillions of events per day with millisecond query response times:

- The serverless architecture eliminates infrastructure management overhead, while automatic data lifecycle management optimizes costs by intelligently tiering data between fast memory storage and cost-effective magnetic storage
- Built-in time series functions and SQL compatibility accelerate development, and seamless AWS service integrations enable comprehensive analytics pipelines
- The service provides enterprise-grade security, compliance certifications, and 99.999999999% data durability with Multi-AZ replication

Service Integration

Amazon Timestream integrates deeply with core AWS services for comprehensive time-series solutions. Data ingestion flows through AWS IoT Core for device data, Amazon Kinesis for streaming analytics, Amazon MSK for Kafka workloads, and AWS Lambda for serverless processing:

- EventBridge Pipes enables real-time data streaming from Kinesis with transformation capabilities
- Analytics integrations include Amazon QuickSight for visualization, Amazon SageMaker for machine learning, Amazon Athena for ad-hoc queries via UNLOAD to S3, and Amazon Managed Grafana for operational dashboards
- Security and management leverage AWS KMS for encryption, IAM for access control, VPC for network isolation, AWS Backup for table backups, and CloudWatch for monitoring with custom metrics and alarms

Onboarding and Offboarding

Customers can quickly start with Amazon Timestream using the AWS Management Console tutorial that creates a database with sample IoT and DevOps datasets, or explore the fully functional sample applications available in multiple programming languages (Java, Python, Node.js, Go, .NET). The getting started process involves creating a database and table through the console, configuring retention policies for memory and magnetic stores, and connecting via AWS SDKs, CLI, or APIs:

- Prerequisites include AWS account setup, IAM permissions configuration, and optionally VPC and security group setup for network isolation
- Sample applications on GitHub demonstrate database creation, data ingestion, and query execution patterns to accelerate development and best practices adoption

Getting Started Guide:
https://docs.aws.amazon.com/timestream/latest/developerguide/getting-started.html

Data Removal

Data removal during offboarding involves deleting individual tables, entire databases, or complete service resources. Tables can be deleted through the console, CLI, or APIs with permanent data removal across all storage tiers. Database deletion removes all contained tables and associated data. For Timestream for InfluxDB clusters, deletion permanently removes all cluster nodes and data stored in InfluxDB Object Storage on Amazon S3 with no recovery options. Critical data should be exported using InfluxDB tools before deletion. The deletion process takes 5-10 minutes for clusters, and billing stops immediately upon completion.

Backup/Restore and Disaster Recovery

Amazon Timestream provides backup capabilities through AWS Backup integration for centralized data protection. Backups are created at table granularity with the first backup being full and subsequent backups incremental. Cross-region and cross-account backup copying enables disaster recovery scenarios. Restore operations can recreate tables in different regions or accounts with configurable retention settings. Multi-AZ data replication provides built-in resilience against zone failures. However, Timestream for InfluxDB lacks native snapshot capabilities, requiring data export for backup purposes.

Backup Capabilities

Timestream for LiveAnalytics integrates with AWS Backup for comprehensive table backup management. Users can create on-demand backups or scheduled backups through backup plans with configurable retention periods and lifecycle policies:

- Incremental backups optimize storage costs while maintaining complete recovery capabilities
- AWS Backup provides centralized policy management, cross-region replication, and automated backup scheduling
- Timestream for InfluxDB does not have native backup capabilities and requires manual data export using InfluxDB tools for data protection

Disaster Recovery

Amazon Timestream supports disaster recovery through multi-region backup copying and restoration capabilities via AWS Backup. Cross-region backup replication enables recovery in different AWS regions during regional outages:

- The service's Multi-AZ architecture provides automatic failover within regions for high availability
- Application Recovery Controller can orchestrate traffic routing during failures
- However, Timestream for InfluxDB has limited disaster recovery options due to lack of native backup features, requiring custom data export and replication strategies

Recovery Objectives

Amazon Timestream helps customers achieve recovery objectives through automated Multi-AZ replication for sub-minute RPO within regions and cross-region backup capabilities for disaster recovery. The serverless architecture enables rapid scaling during recovery scenarios. AWS Backup's incremental backup strategy minimizes recovery time while backup automation reduces RPO. For RTO objectives, the service's cellular architecture and auto-scaling capabilities support fast recovery and capacity restoration. Customers can implement backup and restore strategies with 2-4 hour RTO/RPO for most workloads through proper backup planning and cross-region replication.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/general/latest/gr/timestream.html

FAQ: https://aws.amazon.com/timestream/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/timestream/latest/developerguide/what-is-timestream.html

User Guide: https://docs.aws.amazon.com/timestream/latest/developerguide/getting-started.html

API Reference: https://docs.aws.amazon.com/timestream/latest/developerguide/API_Reference.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/timestream/pricing/

Cost Optimization

Amazon Timestream offers unique cost optimization through intelligent data lifecycle management with automatic tiering between memory and magnetic storage based on data age and access patterns. The Timestream Compute Units (TCU) model provides predictable query costs with per-second billing and configurable MaxQueryTCU limits for budget control:

- Multi-measure records reduce storage costs by storing multiple metrics in single rows
- Batch writes optimize ingestion costs, while query optimization through time predicates, measure name filtering, and spatial pruning reduces compute consumption
- The adaptive query engine automatically optimizes performance and costs, and scheduled queries enable automated data processing to minimize interactive query expenses

Savings Considerations

Primary cost savings come from Timestream's serverless architecture eliminating infrastructure management overhead and providing 10x cost reduction versus relational databases for time-series workloads. Automatic data lifecycle management significantly reduces storage costs by moving older data to cost-optimized magnetic storage while maintaining query access:

- The service's compression and optimized storage format minimize storage footprint
- Pay-per-use pricing for writes, storage, and queries ensures customers only pay for actual consumption without pre-provisioning
- Batch writing reduces per-record costs, while query optimization techniques like limiting scanned data and using time filters minimize compute charges
- AWS Backup's incremental backup strategy reduces backup storage costs compared to full backup alternatives

Service Name

Amazon Transcribe

Service Overview

Amazon Transcribe is a fully managed automatic speech recognition (ASR) service that converts speech to text using advanced machine learning models. It enables developers to add speech-to-text capabilities to their applications by processing both live and recorded audio or video input to produce accurate transcripts. The service supports both real-time streaming transcription and batch processing of audio files, handling multiple speakers, languages, and audio formats while providing features like speaker identification, punctuation, timestamps, and content filtering for comprehensive transcription solutions.

Key Use Cases

- Contact center analytics: Transcribing customer service calls for quality monitoring, compliance, and extracting insights from customer interactions
- Media content captioning: Creating subtitles and closed captions for video content, live streams, and on-demand media to improve accessibility
- Meeting and conference transcription: Converting recorded meetings, webinars, and conference calls into searchable text documents for documentation and analysis
- Voice-enabled applications: Adding speech-to-text capabilities to applications for voice commands, dictation, and conversational interfaces
- Medical documentation: Using Amazon Transcribe Medical for clinical documentation, physician-dictated notes, and healthcare voice applications with HIPAA compliance
- Legal transcription: Converting depositions, court proceedings, and legal recordings into text for case management and documentation

Features

- **Batch and Streaming Transcription**: Supports both batch processing of audio files and real-time streaming transcription with WebSocket and HTTP/2 protocols
- **Multi-language Support**: Supports over 100 languages with automatic language identification for multi-lingual audio streams
- **Speaker Diarization**: Identifies and separates speech from up to 10 different speakers in audio recordings
- **Custom Vocabularies and Language Models**: Allows creation of custom vocabularies and domain-specific language models to improve accuracy for specialized terminology

- **Content Redaction and Filtering**: Automatically identifies and redacts personally identifiable information (PII) and filters toxic content
- **Call Analytics**: Provides detailed call analytics including sentiment analysis, call summarization, and conversation insights for contact centers
- **Medical Transcription**: HIPAA-eligible service variant designed specifically for medical and healthcare transcription use cases

Value Proposition

Amazon Transcribe offers superior accuracy with machine learning models trained on diverse datasets, providing high-quality transcriptions across demographic groups and audio conditions. Unlike traditional transcription services, it delivers real-time processing capabilities, automatic scaling, and pay-as-you-use pricing without upfront costs or long-term contracts:

- The service integrates seamlessly with other AWS services for complete workflow automation, includes built-in security features like PII redaction and encryption, and offers specialized variants for healthcare (HIPAA-compliant) and contact centers
- Its comprehensive feature set includes speaker identification, custom vocabulary support, multi-language capabilities, and advanced analytics, making it more versatile than basic transcription alternatives while maintaining enterprise-grade reliability and compliance standards

Service Integration

Amazon Transcribe integrates deeply with Amazon S3 for storing input audio files and output transcripts, with automatic triggering through S3 events. It works seamlessly with AWS Lambda for serverless processing workflows and event-driven transcription automation:

- Integration with Amazon CloudWatch provides comprehensive monitoring and metrics tracking, while AWS IAM ensures secure access control and permissions management
- The service connects with Amazon EventBridge for workflow orchestration and can trigger downstream processing through Amazon SNS notifications
- For advanced analytics, it integrates with Amazon Bedrock for AI-powered summarization and insights extraction from transcripts, and works with Amazon API Gateway for building transcription APIs and web applications

Onboarding and Offboarding

Getting started with Amazon Transcribe requires an AWS account, IAM credentials configuration, and an Amazon S3 bucket for batch transcriptions. Customers can begin

immediately using the AWS Management Console for simple transcriptions, or install the AWS CLI and SDKs for programmatic access:

- The service offers multiple transcription methods including console interface, CLI commands, REST APIs, and WebSocket streaming
- For beginners, the AWS Management Console provides the easiest entry point with guided interfaces
- Advanced users can leverage SDKs in multiple programming languages (Python, Java, JavaScript, .NET) with comprehensive code examples and documentation
- The service includes a free tier for new customers to explore capabilities before scaling to production workloads

Getting Started Guide: https://docs.aws.amazon.com/transcribe/latest/dg/getting-started.html

Data Removal

Amazon Transcribe provides complete data control through the DeleteTranscriptionJob API and console interface, allowing customers to delete transcription jobs and associated transcripts on-demand. Job records are automatically retained for 90 days before deletion. For offboarding, customers can use batch deletion commands to remove all transcription jobs and associated data. The service does not store voice inputs permanently, and customers maintain full control over their audio files stored in S3 buckets, enabling complete data removal when needed.

Backup/Restore and Disaster Recovery

Amazon Transcribe operates as a fully managed service with AWS handling infrastructure backup and disaster recovery automatically across multiple Availability Zones. Transcript outputs are stored in customer-controlled S3 buckets, inheriting S3's built-in durability and backup capabilities. While the service itself doesn't require backup, customers should implement backup strategies for their source audio files and generated transcripts using S3 versioning, cross-region replication, or AWS Backup integration.

Backup Capabilities

Amazon Transcribe does not have native backup capabilities as it is a processing service rather than a data storage service. The service processes audio and generates transcripts stored in customer S3 buckets. AWS Backup does not directly support Transcribe, but customers can use AWS Backup to protect their S3 buckets containing audio files and transcripts, providing comprehensive data protection for transcription workflows.

Disaster Recovery

Amazon Transcribe supports disaster recovery through AWS's multi-region architecture and automatic failover capabilities. The service is available across multiple AWS regions,

enabling customers to implement cross-region disaster recovery strategies. While it doesn't directly integrate with AWS Elastic Disaster Recovery, customers can architect resilient transcription workflows using multi-region deployments, S3 cross-region replication, and Lambda functions for automated failover scenarios.

Recovery Objectives

Amazon Transcribe helps meet RPO and RTO objectives through its serverless architecture and real-time processing capabilities. The service provides near-zero RPO for streaming transcriptions and supports batch processing for accumulated audio backlogs. With automatic scaling and multi-region availability, customers can achieve low RTO requirements by implementing cross-region failover strategies and maintaining backup processing capabilities in secondary regions.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/general/latest/gr/transcribe.html

FAQ: https://aws.amazon.com/transcribe/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/transcribe/latest/dg/what-is.html

User Guide: https://docs.aws.amazon.com/transcribe/latest/dg/getting-started.html

API Reference: https://docs.aws.amazon.com/transcribe/latest/APIReference/

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/transcribe/pricing/

Cost Optimization

Amazon Transcribe offers several cost optimization strategies including pay-per-second billing with no upfront costs or minimum commitments, reducing waste compared to fixed-price alternatives. Customers can optimize costs by using appropriate audio quality settings - lower sample rates for voice-only content and higher rates only when necessary:

- Batch processing is more cost-effective than streaming for non-real-time requirements
- Volume discounts are available for large workloads, and the free tier provides cost savings for initial usage
- Custom language models and vocabulary filtering can reduce processing time and improve accuracy, potentially reducing re-processing costs

Savings Considerations

Main cost savings come from eliminating expensive manual transcription services that can cost 60% more than automated solutions. The service eliminates infrastructure management overhead and staffing costs for transcription operations:

- Pay-as-you-use pricing scales with actual usage, avoiding over-provisioning costs
- Integration with other AWS services creates workflow efficiencies that reduce operational expenses
- Job queueing capabilities allow efficient resource utilization during peak periods
- The service's high accuracy reduces post-processing and editing costs, while automated features like speaker identification and content filtering eliminate manual review requirements for many use cases

Service Name

Amazon Transcribe Medical

Service Overview

Amazon Transcribe Medical is an automatic speech recognition (ASR) service specifically designed for medical professionals to accurately transcribe medical-related speech into text. The service converts physician-dictated notes, drug safety monitoring calls, telemedicine appointments, and physician-patient conversations into accurate text transcriptions. It is HIPAA-eligible and designed to handle medical terminology across multiple specialties including cardiology, neurology, oncology, primary care, radiology, and urology. The service supports both real-time streaming transcription and batch processing of audio files, making it suitable for various healthcare applications from clinical documentation to pharmacovigilance workflows.

Key Use Cases

- Clinical documentation: Automatically transcribe physician-patient conversations and clinician-dictated notes for electronic health record (EHR) systems
- Drug safety monitoring (pharmacovigilance): Generate text transcripts from recorded calls to identify reported side effects and adverse events
- Telehealth and telemedicine: Provide real-time speech recognition during remote consultations and virtual appointments
- Medical dictation applications: Convert spoken medical reports and patient visit summaries into structured text
- Contact center analytics: Analyze healthcare-related customer service calls and support interactions

Features

- **Medical specialty support**: Supports transcription for cardiology, neurology, oncology, primary care (family medicine, internal medicine, OB-GYN, pediatrics), radiology, and urology specialties
- **Real-time and batch processing**: Offers both streaming transcription for live conversations and batch transcription for audio files
- **Medical terminology accuracy**: Accurately transcribes medical terms and converts measurements into standard abbreviations (e.g., blood pressure to BP)
- **Speaker identification**: Enables speaker partitioning to identify different speakers in medical conversations
- **Custom medical vocabularies**: Supports custom vocabularies to improve transcription accuracy for specific medical terminology

- **Alternative transcriptions**: Provides multiple transcription alternatives with confidence scores
- **Multi-channel audio support**: Processes multi-channel audio recordings for comprehensive conversation capture

Value Proposition

Amazon Transcribe Medical offers healthcare organizations a purpose-built solution that understands medical terminology and context, delivering superior accuracy compared to generic speech recognition services. The service is HIPAA-eligible and designed with healthcare compliance requirements in mind, ensuring patient data privacy and security:

- It reduces development time and total cost of ownership by providing ready-to-use APIs that eliminate the need for in-house ASR expertise
- The service integrates seamlessly with existing healthcare workflows and can be combined with other AWS services like Amazon Comprehend Medical for enhanced text analysis
- Its specialty-specific models and support for medical custom vocabularies ensure high-quality transcriptions tailored to different medical domains

Service Integration

Amazon Transcribe Medical integrates with several core AWS services to create comprehensive healthcare solutions. It stores transcription outputs in Amazon S3 buckets and can be triggered by AWS Lambda functions for automated processing workflows:

- The service works with Amazon Comprehend Medical for advanced medical text analysis and entity extraction from transcribed content
- Integration with Amazon Connect enables healthcare contact center analytics, while Amazon Chime SDK integration supports telehealth applications
- The service can be combined with Amazon Translate for multilingual capabilities and Amazon Polly for text-to-speech functionality
- CloudWatch provides monitoring and logging capabilities, and IAM controls access and security permissions for healthcare compliance requirements

Onboarding and Offboarding

Customers can start with Amazon Transcribe Medical through the AWS Management Console or by using AWS SDKs and APIs. The service offers a free tier for initial testing and evaluation:

- Getting started requires creating transcription jobs through the console, specifying audio file locations, medical specialties, and transcription types (dictation or conversation)

- For real-time applications, customers can use streaming APIs with WebSocket or HTTP/2 protocols
- The service requires minimal setup - customers need to configure IAM permissions, specify medical specialties, and choose between batch or streaming processing
- No machine learning expertise is required, as the service handles all ASR processing in the backend

Getting Started Guide: https://docs.aws.amazon.com/transcribe/latest/dg/transcribe-medical.html

Data Removal

Customers can delete medical transcription jobs and their associated transcripts using the DeleteMedicalTranscriptionJob API operation. The service does not store customer voice inputs or use them for model training. Transcription outputs stored in customer-owned S3 buckets can be deleted by customers at any time. Amazon Transcribe Medical operates under AWS's shared responsibility model, where customers are responsible for managing their data lifecycle and deletion policies.

Backup/Restore and Disaster Recovery

Amazon Transcribe Medical does not store customer voice inputs and only retains job records for 90 days. The service relies on AWS infrastructure for high availability and disaster recovery. Customers are responsible for backing up transcription outputs stored in their S3 buckets using standard S3 backup and versioning features. The service itself does not provide built-in backup capabilities as it operates as a stateless transcription service.

Backup Capabilities

Amazon Transcribe Medical does not have built-in backup capabilities as it is a stateless processing service. The service does not store customer audio data and only retains job records for 90 days:

- Customers must implement their own backup strategies for transcription outputs stored in S3 buckets
- The service does not integrate with AWS Backup service directly

Disaster Recovery

Amazon Transcribe Medical inherits disaster recovery capabilities from the underlying AWS infrastructure but does not directly integrate with AWS Elastic Disaster Recovery. The service is available across multiple AWS regions for geographic redundancy:

- Since the service is stateless and doesn't store customer data, disaster recovery focuses on service availability rather than data recovery

- Customers can implement cross-region strategies using multiple AWS regions

Recovery Objectives

Amazon Transcribe Medical helps customers achieve low RTO objectives through its stateless architecture and multi-region availability. Since the service doesn't store customer audio data, RPO is primarily determined by how frequently customers back up their transcription outputs in S3. The service's real-time streaming capabilities support near-zero RTO for live transcription scenarios, while batch processing can resume quickly after any service interruption.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/general/latest/gr/transcribe.html#limits-amazon-transcribe

FAQ: https://aws.amazon.com/transcribe/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/transcribe/latest/dg/transcribe-medical.html

User Guide: https://docs.aws.amazon.com/transcribe/latest/dg/transcribe-medical.html

API Reference: https://docs.aws.amazon.com/transcribe/latest/api/

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/transcribe/pricing/

Cost Optimization

Amazon Transcribe Medical offers several cost optimization strategies unique to healthcare transcription workloads. The service charges only for actual audio duration transcribed, not per API call, making it cost-effective for variable-length medical dictations:

- Batch processing is more cost-effective than streaming for non-real-time use cases
- Customers can optimize costs by preprocessing audio to remove silence and using appropriate audio formats to reduce file sizes
- The service's specialty-specific models reduce the need for expensive post-processing corrections
- Integration with other AWS services through bulk processing can achieve volume discounts, and the free tier allows for cost-effective testing and small-scale deployments

Savings Considerations

The main cost savings considerations for Amazon Transcribe Medical include replacing expensive human transcription services with automated processing, reducing the total cost of ownership through elimination of infrastructure management and ASR expertise requirements. The service eliminates upfront capital expenses for speech recognition software and hardware:

- Healthcare organizations can achieve significant labor cost reductions while improving transcription turnaround times
- The pay-per-use model prevents over-provisioning costs, and the service's high accuracy reduces expensive correction and editing workflows
- Integration with existing AWS healthcare solutions can create economies of scale, and the HIPAA-eligible status eliminates costs associated with additional compliance infrastructure and certifications

Service Name

Amazon Virtual Private Cloud

Service Overview

Amazon Virtual Private Cloud (VPC) enables users to launch AWS resources in a logically isolated virtual network within the AWS cloud. It provides complete control over networking environment including IP address ranges, subnets, routing tables, internet gateways, and security settings. VPC resembles a traditional network but operates in the cloud, allowing organizations to define their own virtual network topology with customizable security policies, access controls, and connectivity options. The service supports both IPv4 and IPv6 addressing, provides robust logical isolation with individually authorized packet flows, and includes features like security groups, network ACLs, VPC Flow Logs, peering connections, and VPN connectivity for comprehensive network management.

Key Use Cases

- Web and database servers: Creating production environments with web servers in public subnets and database servers in private subnets for multi-tier applications
- Private server environments: Deploying all EC2 instances in private subnets with NAT gateways for secure outbound internet access without direct inbound connectivity
- Test and development environments: Setting up isolated development or testing environments with customizable network configurations
- Hybrid cloud connectivity: Connecting on-premises networks to AWS via VPN or Direct Connect for seamless hybrid architectures
- Multi-account and multi-region architectures: Using VPC peering and Transit Gateway to connect multiple VPCs across accounts and regions for enterprise-scale deployments

Features

- **Subnets and IP Addressing**: Create public and private subnets with customizable IPv4 and IPv6 CIDR blocks, supporting up to 5 CIDR blocks per VPC with expansion to 50
- **Security Groups and Network ACLs**: Instance-level and subnet-level security controls with stateful and stateless filtering for granular traffic management
- **Internet and NAT Gateways**: Provide internet connectivity for public subnets and secure outbound access for private subnets through managed NAT services
- **VPC Peering and Transit Gateway**: Connect multiple VPCs within same region or across regions for scalable network architectures with centralized routing

- **VPC Flow Logs**: Monitor and log IP traffic going to and from network interfaces for security analysis and troubleshooting
- **AWS PrivateLink and VPC Endpoints**: Access AWS services privately without internet gateways, maintaining traffic within AWS network backbone
- **DNS Resolution**: Built-in DNS services with customizable DHCP option sets for name resolution within VPC
- **Route Tables**: Define network traffic routing with up to 200 route tables per VPC and 500 non-propagated routes per table

Value Proposition

Amazon VPC provides unparalleled control and security for cloud networking with logical isolation that makes unauthorized information transfer highly improbable. Unlike other cloud networking solutions, VPC offers complete customization of network topology, IP addressing, routing, and security policies while maintaining seamless integration with all AWS services:

- The service provides robust security through multiple layers including security groups, network ACLs, and VPC Flow Logs for comprehensive monitoring
- VPC eliminates the complexity of traditional network hardware management while offering enterprise-grade features like encryption controls, traffic mirroring, and advanced connectivity options
- With no additional charges for basic VPC usage and built-in high availability across multiple Availability Zones, it delivers cost-effective networking that scales automatically with business needs

Service Integration

Amazon VPC serves as the foundational networking layer for virtually all AWS services, providing deep integration across the entire AWS ecosystem. Core compute services like EC2, Lambda, and ECS launch directly into VPC subnets with configurable security groups:

- Database services including RDS, Aurora, DynamoDB, and ElastiCache operate within VPC for secure data access
- VPC integrates seamlessly with AWS PrivateLink for private service connectivity, Route 53 for DNS services, and CloudFront for content delivery
- Storage services like S3 and EFS connect through VPC endpoints, while monitoring services such as CloudWatch, CloudTrail, and VPC Flow Logs provide comprehensive visibility
- Additional integrations include AWS Network Firewall for advanced security, Direct Connect for hybrid connectivity, and Transit Gateway for multi-VPC architectures, creating a cohesive and secure cloud infrastructure foundation

Onboarding and Offboarding

Getting started with Amazon VPC is straightforward through multiple pathways designed for different user preferences and expertise levels. AWS provides a default VPC in each region that's immediately ready for use, allowing users to launch resources without additional configuration:

- For custom requirements, users can create new VPCs through the AWS Management Console with guided wizards that automatically configure subnets, gateways, and routing tables
- The AWS CLI and SDKs offer programmatic access for infrastructure-as-code deployments
- AWS provides comprehensive tutorials covering basic public subnet configurations for simple web applications and advanced multi-tier architectures with private subnets and NAT gateways
- Getting started guides include hands-on examples for creating VPCs with EC2 instances, connecting to on-premises networks, and implementing security best practices
- CloudFormation templates are available for standardized deployments across multiple environments

Getting Started Guide: https://docs.aws.amazon.com/vpc/latest/userguide/getting-started-with-amazon-vpc-using-the-aws-cli.html

Data Removal

Offboarding from Amazon VPC involves systematically removing resources in the correct order to avoid dependency conflicts. Users must first terminate or delete all resources within the VPC including EC2 instances, RDS databases, load balancers, and NAT gateways. Network interfaces, security groups, and route table associations must be cleaned up before removing subnets. VPC peering connections, VPN connections, and internet gateways must be detached and deleted. Finally, the VPC itself can be deleted once all dependencies are removed. AWS provides detailed cleanup procedures and automated tools to identify and remove VPC resources systematically.

Backup/Restore and Disaster Recovery

Amazon VPC itself is a control plane service where the VPC configuration is automatically maintained and replicated by AWS across multiple Availability Zones within a region. VPC configurations including subnets, route tables, security groups, and network ACLs are automatically backed up as part of AWS service management. For disaster recovery, VPC configurations can be recreated in other regions using Infrastructure as Code tools like CloudFormation or Terraform. Cross-region VPC peering and Transit Gateway connections enable multi-region architectures for disaster recovery scenarios.

Backup Capabilities

VPC configurations are automatically maintained and replicated by AWS as part of the managed service infrastructure. The service integrates with AWS Backup for protecting resources within the VPC such as EC2 instances, EBS volumes, and RDS databases:

- VPC Flow Logs can be stored in S3 or CloudWatch Logs with versioning and lifecycle policies for long-term retention
- Infrastructure as Code templates provide version-controlled backup of VPC configurations that can be deployed across regions for disaster recovery purposes

Disaster Recovery

Amazon VPC supports comprehensive disaster recovery strategies through multi-region deployment capabilities and integration with AWS disaster recovery services. VPC configurations can be replicated across regions using Infrastructure as Code, while cross-region peering and Transit Gateway enable multi-region connectivity:

- The service integrates with AWS Application Recovery Controller for automated failover orchestration
- Resources within VPCs can leverage AWS Elastic Disaster Recovery for continuous replication and rapid recovery of applications and databases to alternate regions or Availability Zones

Recovery Objectives

Amazon VPC enables customers to achieve aggressive RPO and RTO objectives through multi-Availability Zone deployments that provide automatic failover within minutes. Cross-region VPC architectures with automated failover can achieve RPO of near-zero and RTO of minutes to hours depending on application complexity. VPC's integration with AWS services like Route 53 health checks, Application Load Balancer failover, and RDS Multi-AZ deployments supports sub-minute RTO for many scenarios. The service's ability to maintain persistent IP addresses and network configurations across Availability Zones minimizes recovery complexity and time.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/vpc/latest/userguide/amazon-vpc-limits.html

FAQ: https://aws.amazon.com/vpc/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/vpc/latest/userguide/what-is-amazon-vpc.html

User Guide: https://docs.aws.amazon.com/vpc/latest/userguide/what-is-amazon-vpc.html

API Reference: https://docs.aws.amazon.com/AWSEC2/latest/APIReference/OperationList-query-vpc.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/vpc/pricing/

Cost Optimization

Amazon VPC offers exceptional cost optimization through its core service being free of charge, with costs only for additional components like NAT gateways, Elastic IP addresses, and data transfer. Customers can optimize costs by using VPC endpoints instead of NAT gateways for AWS service access, reducing data transfer charges:

- Strategic placement of resources within the same Availability Zone eliminates cross-AZ data transfer fees
- IPv6 adoption can reduce reliance on costly public IPv4 addresses
- VPC peering provides cost-effective connectivity compared to Transit Gateway for simple multi-VPC scenarios
- Careful sizing of NAT gateways and strategic use of public subnets for internet-facing resources minimize ongoing operational costs while maintaining security and performance requirements

Savings Considerations

Primary cost savings come from VPC's core functionality being provided at no charge, eliminating traditional networking hardware and maintenance costs. Customers save significantly by using VPC endpoints for AWS service access instead of routing through NAT gateways, reducing both data processing and transfer charges:

- Designing efficient network topologies that minimize cross-region and cross-AZ traffic reduces data transfer costs
- Implementing shared VPCs across multiple AWS accounts eliminates redundant networking infrastructure
- Using security groups and network ACLs instead of third-party security appliances provides enterprise-grade security without additional licensing costs
- Strategic use of Availability Zone placement and careful planning of public versus private subnet usage optimizes both cost and security

- Reserved capacity for predictable workloads and rightsizing NAT gateway instances based on actual throughput requirements further optimize ongoing operational expenses

Service Name

Amazon WorkSpaces

Service Overview

Amazon WorkSpaces is a fully managed virtual desktop infrastructure (VDI) service that enables organizations to provision virtual, cloud-based desktops for users. These desktops can run Microsoft Windows, Amazon Linux, Ubuntu Linux, Rocky Linux, or Red Hat Enterprise Linux. WorkSpaces eliminates the need to procure and deploy hardware or install complex software, allowing users to access their virtual desktops from multiple devices or web browsers. The service offers two options: WorkSpaces Personal for persistent virtual desktops tailored for users requiring highly-personalized desktop environments, and WorkSpaces Pools for non-persistent virtual desktops for users needing access to curated desktop environments on ephemeral infrastructure.

Key Use Cases

- Remote and hybrid work: Enable secure access to corporate applications and data from anywhere
- Learning environments: Provide virtual desktops for educational institutions and training programs
- Developer workstations: On-demand access to high-performance development environments
- Contact centers: Secure virtual desktops for customer service representatives
- Knowledge workers: Consistent desktop experience for office productivity applications
- Running Windows 11 on older devices: Extend hardware lifecycle by accessing modern OS through virtual desktops
- Back office and administration: Secure environments for administrative and financial operations
- Secure development workflows: Highly regulated environments meeting compliance requirements

Features

- **Multi-Platform Access**: Access WorkSpaces from Windows, Mac, Linux, Chromebooks, iPads, Android devices, and web browsers
- **Flexible Hardware Bundles**: Multiple compute types from Value to GraphicsPro.g4dn with varying vCPU, memory, and storage configurations
- **Multiple Operating Systems**: Support for Windows, Amazon Linux, Ubuntu Linux, Rocky Linux, and Red Hat Enterprise Linux

- **Custom Images and Bundles**: Create and deploy custom desktop images with pre-installed applications and configurations
- **Multi-Factor Authentication**: Integration with RADIUS servers and SAML 2.0 identity providers for enhanced security
- **Data Encryption**: Encryption at rest and in transit using AWS Key Management Service
- **Persistent Storage**: Automatic backup of user data to Amazon S3 with configurable storage sizes
- **Streaming Protocols**: Support for WorkSpaces Streaming Protocol (WSP) and PCoIP for optimized performance
- **Self-Service Management**: Users can restart, rebuild, change compute type, and manage their WorkSpaces directly
- **Multi-Region Resilience**: Cross-region redirection and standby WorkSpaces for disaster recovery

Value Proposition

Amazon WorkSpaces delivers significant advantages over traditional VDI solutions through its fully managed cloud infrastructure that eliminates hardware procurement, complex software installation, and ongoing maintenance overhead. Organizations benefit from rapid deployment capabilities, with virtual desktop deployment times reduced by up to 90% as demonstrated by companies like Ferrari:

- The service provides unmatched flexibility with support for both persistent and non-persistent desktops, multiple operating systems, and global scalability across AWS regions
- Cost optimization is built-in through flexible pricing models including AlwaysOn and AutoStop billing, integration with EC2 Reserved Instances and Savings Plans, and automated cost optimization tools
- Enhanced security features include encryption, multi-factor authentication, and Zero Trust architecture compatibility, while seamless integration with Microsoft 365 and existing enterprise systems ensures smooth adoption

Service Integration

Amazon WorkSpaces integrates extensively with core AWS services to deliver a comprehensive virtual desktop solution. Directory integration is provided through AWS Directory Service with support for Simple AD, AWS Managed Microsoft AD, and Active Directory Connector for seamless user authentication:

- Security is enhanced through AWS Key Management Service for encryption, AWS IAM for access management, and integration with RADIUS servers for multi-factor authentication

- The service leverages Amazon VPC for network isolation and security groups for access control, while Amazon CloudWatch provides comprehensive monitoring and logging capabilities
- Storage integration includes automatic backup to Amazon S3 and compatibility with Amazon FSx for Windows File Server
- For disaster recovery, WorkSpaces utilizes cross-region capabilities and integrates with AWS Backup for data protection, while AWS Systems Manager enables automated patching and configuration management

Onboarding and Offboarding

Getting started with Amazon WorkSpaces requires three essential components: a compatible client application, a directory service for user authentication, and an Amazon VPC with at least two subnets. Organizations first create a directory using Simple AD, AWS Managed Microsoft AD, or Active Directory Connector, then configure the VPC with proper security groups and DHCP options:

- WorkSpaces can be launched through the AWS Management Console by selecting bundle types, running modes, and directory settings
- Users receive invitation emails with client download instructions and connection details
- The setup process includes configuring multi-factor authentication if required, customizing bundles and images as needed, and establishing network connectivity
- Best practices recommend deploying WorkSpaces in private subnets with proper security group rules for Active Directory communications and implementing monitoring through Amazon CloudWatch for ongoing management

Getting Started Guide:
https://docs.aws.amazon.com/workspaces/latest/adminguide/amazon-workspaces.html

Data Removal

Offboarding from Amazon WorkSpaces involves terminating WorkSpaces through the AWS Management Console, CLI, or API, which permanently deletes all data stored on root and user volumes. Organizations should backup critical data before termination since the process is irreversible. Directory services and associated infrastructure components like VPCs and security groups must be separately deregistered and deleted. For compliance requirements, encryption keys used for data protection should be managed according to organizational policies, and any custom images or bundles can be deleted to remove organizational data completely.

Backup/Restore and Disaster Recovery

Amazon WorkSpaces provides comprehensive backup and disaster recovery capabilities through automatic snapshots of root and user volumes stored in Amazon S3. Multi-Region

Resilience (MRR) enables creation of standby WorkSpaces in secondary regions with cross-region redirection and DNS health check failover. Data replication can copy user data one-way from primary to secondary regions. WorkSpaces can be restored to previous snapshots when in AVAILABLE, ERROR, UNHEALTHY, or STOPPED states, though this overwrites current data with snapshot contents.

Backup Capabilities

WorkSpaces automatically creates snapshots of both root and user volumes after initial creation, during regular use, and after restoration events. These snapshots are stored in Amazon S3 for durability and can be used to restore WorkSpaces to previous states:

- The service integrates with AWS Backup for centralized backup management across AWS services
- However, WorkSpaces snapshots are service-specific and restoration requires both root and user volume snapshots to be available

Disaster Recovery

WorkSpaces supports disaster recovery through Multi-Region Resilience, enabling standby WorkSpaces in secondary AWS regions with automated failover capabilities. Cross-region redirection uses DNS health checks for automatic traffic routing during outages. While WorkSpaces doesn't directly integrate with AWS Elastic Disaster Recovery, organizations can implement comprehensive DR strategies using AWS Backup for data protection and multi-region deployment architectures for business continuity.

Recovery Objectives

Amazon WorkSpaces helps customers meet Recovery Point Objective (RPO) and Recovery Time Objective (RTO) requirements through automated snapshot creation and Multi-Region Resilience features. Standby WorkSpaces can be activated within minutes during regional failures, supporting aggressive RTO targets. Data replication capabilities enable near real-time data synchronization for minimal RPO requirements. The fully managed infrastructure and automated failover mechanisms significantly reduce recovery complexity while providing predictable recovery timeframes for business continuity planning.

Service Constraints

Service Quotas:
https://docs.aws.amazon.com/workspaces/latest/adminguide/workspaces-limits.html

FAQ: https://aws.amazon.com/workspaces/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/workspaces/latest/adminguide/

User Guide: https://docs.aws.amazon.com/workspaces/latest/userguide/

API Reference: https://docs.aws.amazon.com/workspaces/latest/api/

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/workspaces/pricing/

Cost Optimization

Amazon WorkSpaces offers unique cost optimization features including the Cost Optimizer solution that automatically analyzes usage patterns and converts WorkSpaces between hourly and monthly billing models to minimize costs. AutoStop mode automatically stops WorkSpaces when not in use, charging only for active hours plus a small monthly fee:

- Organizations can leverage EC2 Reserved Instances and Savings Plans for additional savings, while the ability to terminate unused WorkSpaces and optimize pool sizes reduces infrastructure costs
- Bring Your Own License (BYOL) options for Windows and Microsoft Office licenses provide significant cost savings compared to license-included bundles
- The service eliminates traditional VDI infrastructure costs including hardware procurement, maintenance, and data center operations

Savings Considerations

Primary cost savings with Amazon WorkSpaces come from eliminating upfront hardware investments and ongoing infrastructure maintenance costs associated with traditional VDI deployments. The pay-as-you-use model with AlwaysOn and AutoStop billing options ensures organizations only pay for actual usage rather than overprovisioned capacity:

- Integration with EC2 Reserved Instances and Compute Savings Plans can provide up to 72% cost savings for predictable workloads
- BYOL licensing reduces per-user costs significantly, while automated cost optimization tools prevent bill shock through intelligent workload management
- The fully managed service model eliminates IT overhead for patching, monitoring, and maintenance, allowing organizations to redirect resources to strategic initiatives while reducing total cost of ownership compared to on-premises VDI solutions

Service Name

Amazon WorkSpaces Applications

Service Overview

Amazon WorkSpaces Applications is a fully managed application streaming service that enables organizations to centrally manage and securely deliver desktop applications to any device without requiring code changes. The service transforms desktop applications into cloud-based SaaS solutions, allowing users to access applications through a web browser or native client with a responsive, fluid user experience indistinguishable from locally installed applications. Applications run on AWS compute resources with automatic scaling, pay-as-you-go pricing, and global availability, while data never leaves AWS, ensuring high performance and security.

Key Use Cases

- Converting desktop applications to SaaS: Transform existing desktop applications into cloud-based services without code modifications for software vendors and enterprises
- Learning environments: Provide students and educators with on-demand access to specialized software and applications for educational institutions
- Remote and hybrid workforce enablement: Deliver applications to distributed teams working from various locations and devices
- Application modernization: Migrate legacy applications to the cloud while maintaining familiar user experiences
- Temporary or contractor access: Provide secure application access to temporary workers without local installation requirements

Features

- **Multi-device access**: Access applications from Windows, Linux PCs, Macs, Chromebooks, iPads, and Android tablets via HTML5 browsers or native clients
- **Three fleet types**: Always-On (constant pool), On-Demand (provision on demand), and Elastic (charge only during streaming sessions)
- **Multi-session support**: Multiple user sessions on a single fleet instance for Windows fleets to optimize resource usage and costs
- **Native client features**: Multi-monitor support, 4K resolution, USB device usage, local drive access, print redirection, and webcam support
- **Authentication integration**: Support for User Pool, SAML 2.0 SSO, Microsoft Active Directory, and API-based streaming URLs
- **Auto scaling**: Schedule-based and usage-based scaling policies with configurable minimum and maximum capacity settings

- **Real-time audio-video**: Local webcam and microphone input support for video conferencing within streaming sessions

Value Proposition

Amazon WorkSpaces Applications eliminates the need for infrastructure procurement and maintenance while providing instant global scalability and pay-as-you-go pricing with no upfront investment. Organizations can transform desktop applications into SaaS solutions without code changes, reducing time to market and operational complexity:

- The service ensures high performance with applications running on AWS compute resources, automatic scaling, and data remaining secure in the cloud
- Multi-session capabilities optimize costs by supporting multiple users per instance, while comprehensive authentication integration and enterprise-grade security controls provide seamless user management and compliance adherence

Service Integration

Amazon WorkSpaces Applications integrates deeply with core AWS services including Amazon S3 for application storage and home folder persistence, AWS Identity and Access Management (IAM) for access control and fleet instance roles, and Amazon Elastic Load Balancing for high availability and scalability. The service connects with Amazon VPC for network isolation and security, AWS Directory Service for Microsoft Active Directory integration, and Amazon CloudWatch for monitoring and logging. Additional integrations include Amazon EC2 for underlying compute infrastructure, AWS Lambda for automation workflows, and Amazon FSx for Windows File Server for enterprise file sharing requirements.

Onboarding and Offboarding

Customers can start with the 'Try it Now' feature for a free trial experience or follow the Getting Started tutorial to set up application streaming. The process involves creating a stack (user access policies and storage configurations), choosing or creating an image with installed applications, and configuring a fleet with appropriate instance types and scaling settings:

- Users can be managed through the WorkSpaces Applications user pool, SAML 2.0 SSO, or API-based streaming URLs
- The service automatically creates necessary IAM roles and provides step-by-step guidance through the AWS console for initial setup and configuration

Getting Started Guide:
https://docs.aws.amazon.com/appstream2/latest/developerguide/what-is-how-to-start.html

Data Removal

When offboarding from WorkSpaces Applications, customers can delete stacks, fleets, and images through the AWS console or APIs, which automatically terminates all streaming instances and removes associated resources. Application data stored in home folders or application settings persistence can be removed by deleting the associated Amazon S3 buckets. Custom images and app blocks are deleted from the service, and any stored snapshots or backups should be manually removed from connected storage services.

Backup/Restore and Disaster Recovery

WorkSpaces Applications provides limited native backup capabilities, primarily through image snapshots and application settings persistence to Amazon S3. The service does not include built-in disaster recovery features, but customers can implement cross-region resilience by replicating images, app blocks, and configurations to multiple AWS regions. Organizations must design their own backup strategies for application data using Amazon S3 cross-region replication and implement disaster recovery procedures through Infrastructure as Code deployment in alternate regions.

Backup Capabilities

WorkSpaces Applications does not natively integrate with AWS Backup service. The service primarily relies on image snapshots for application environments and Amazon S3 for persistent data storage like home folders and application settings. Customers must implement custom backup solutions using S3 cross-region replication for data persistence and maintain image catalogs across regions for application availability.

Disaster Recovery

WorkSpaces Applications does not integrate with AWS Elastic Disaster Recovery service. The service lacks built-in multi-region disaster recovery capabilities, requiring customers to manually replicate images, app blocks, stacks, and fleet configurations across regions. Organizations must implement custom disaster recovery procedures using Infrastructure as Code tools like AWS CloudFormation or AWS CDK to recreate environments in alternate regions during outages.

Recovery Objectives

WorkSpaces Applications can help customers achieve recovery objectives through manual cross-region replication of images and configurations, potentially supporting RTO of hours depending on implementation complexity. RPO objectives depend on the frequency of image updates and S3 replication schedules for persistent data. The service’s stateless nature for applications can facilitate faster recovery, but lacks automated failover mechanisms requiring manual intervention for disaster scenarios.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/appstream2/latest/developerguide/limits.html

FAQ: https://aws.amazon.com/workspaces-family/workspaces/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/appstream2/latest/developerguide/what-is-appstream.html

User Guide: https://docs.aws.amazon.com/appstream2/latest/developerguide/getting-started.html

API Reference: https://docs.aws.amazon.com/appstream2/latest/APIReference/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/workspaces/applications/pricing/

Cost Optimization

WorkSpaces Applications offers multiple cost optimization strategies including three fleet types with different pricing models: Always-On for consistent usage, On-Demand for variable usage patterns, and Elastic fleets that charge only during active streaming sessions. Multi-session capabilities allow multiple users per Windows instance, significantly reducing per-user costs:

- The service provides a Simple Pricing Tool and Scaling Optimizer Tool to help estimate costs and optimize fleet capacity
- Organizations can leverage different instance types based on application requirements and implement auto-scaling policies to match capacity with actual demand, avoiding over-provisioning costs

Savings Considerations

Primary cost savings come from eliminating infrastructure procurement, maintenance, and management overhead while paying only for actual usage. Multi-session fleets can reduce costs by 50% or more by sharing instances among users:

- On-Demand and Elastic fleet types provide significant savings for variable usage patterns compared to Always-On fleets

- Educational institutions qualify for reduced Microsoft RDS SAL user fees, and organizations with Microsoft License Mobility can bring their own licenses
- The pay-as-you-go model eliminates upfront capital expenditure and enables precise cost allocation based on actual application usage rather than peak capacity planning

Service Name

Amazon WorkSpaces Secure Browser

Service Overview

Amazon WorkSpaces Secure Browser is a fully managed, cloud-native, hosted browser service that enables secure access to private websites, software-as-a-service (SaaS) applications, and the public internet from a disposable container. Web content is streamed to users' existing web browsers while the actual browser session remains isolated in AWS. Each session launches with a fresh, fully patched Chrome browser, eliminating the need for IT to manage appliances, infrastructure, specialized client software, or VPN connections. The service reduces data exfiltration risk by streaming encrypted pixels without transmitting HTML, DOM, or sensitive company data to local devices.

Key Use Cases

- Empowering customer support and back-office workers: Secure access to internal applications and customer data without exposing sensitive information to local devices
- Enabling BYOD (Bring Your Own Device) initiatives: Allow employees to safely access company resources from personal devices without security risks
- Creating secure data and analytics environments: Provide controlled access to sensitive data and analytics tools while maintaining data isolation
- Isolating high-security networks from the public internet: Enable secure browsing for users in high-security environments without direct internet exposure
- Cost optimization: Replace expensive VPN infrastructure and traditional virtual desktops with a more cost-effective browser solution

Features

- **Ephemeral Sessions**: Each session launches with a fresh, fully patched Chrome browser that is wiped clean after use
- **Pixel Streaming**: Streams encrypted pixels from remote browser sessions without transmitting sensitive data to local devices
- **Enterprise Browser Policies**: Enforce URL allow/blocking, session-level controls for clipboard, file transfer, and printer access
- **IP Access Controls**: Restrict access to trusted networks or devices through IP-based access controls
- **SAML 2.0 Integration**: Single sign-on support with existing SAML 2.0 identity providers for authentication

- **Web Content Filtering**: Control and monitor web content access with granular policies and 25+ predefined categories
- **Session Monitoring**: Monitor sessions using CloudWatch and CloudTrail with detailed logging capabilities
- **Multiple Instance Types**: Support for standard.regular, standard.large, and standard.xlarge instances for different performance needs

Value Proposition

Amazon WorkSpaces Secure Browser offers significant advantages over traditional VPN solutions and virtual desktops. It eliminates the browser attack surface by isolating sessions in AWS, reduces data exfiltration risks through pixel streaming, and requires no specialized client software or VPN management:

- Starting at $7 per month, it provides cost-effective secure browsing compared to traditional virtual desktops
- The service simplifies IT management with fully managed infrastructure, automatic browser updates, and centralized policy enforcement
- Organizations benefit from enhanced security controls, simplified BYOD enablement, and the ability to provide secure access from any device without complex infrastructure management

Service Integration

WorkSpaces Secure Browser integrates deeply with core AWS services to provide secure, scalable browsing capabilities. It leverages Amazon EC2 for provisioning streaming instances and creating Cross-Account Elastic Network Interfaces for corporate network communication:

- The service uses Amazon VPC for network isolation, subnets, and security groups configuration
- Data protection is ensured through integration with AWS Key Management Service (KMS) for encryption, Amazon DynamoDB and S3 for secure data storage, and AWS Identity and Access Management (IAM) for access control
- Monitoring capabilities are provided through Amazon CloudWatch for metrics and AWS CloudTrail for audit logging
- Optional integration with Kinesis Data Streams enables advanced user access logging

Onboarding and Offboarding

Getting started with WorkSpaces Secure Browser involves creating a web portal in the desired AWS region, configuring VPC, subnets, and security groups for network connectivity. Administrators set up browser policies and user settings through the AWS

console, then federate their existing SAML 2.0 identity provider for user authentication and single sign-on:

- The setup process includes signing up for an AWS account, creating IAM users with appropriate permissions, and configuring networking prerequisites
- Once configured, users can access the browser through common web browsers from desktop, laptop, or thin client devices without requiring additional software installation

Getting Started Guide: https://docs.aws.amazon.com/workspaces-web/latest/adminguide/setting-up.html

Data Removal

WorkSpaces Secure Browser ensures data protection through ephemeral sessions that are automatically wiped at the end of each session. No HTML, DOM, or sensitive company data is transmitted to or stored on local devices. When offboarding, administrators can simply deactivate user access through their SAML identity provider and delete web portals through the AWS console. All session data, browser policies, and user preferences stored in AWS are encrypted and can be removed by deleting the associated resources.

Backup/Restore and Disaster Recovery

WorkSpaces Secure Browser provides inherent data protection through its ephemeral session model where no persistent data is stored in browser instances. The service automatically maintains fresh, fully patched browser sessions that are recreated for each use. Configuration data including portal settings, browser policies, and user preferences are stored in AWS managed services with built-in durability and availability features across multiple availability zones.

Backup Capabilities

The service does not require traditional backup capabilities as browser sessions are ephemeral and automatically wiped after use. Configuration data is stored in AWS managed services like DynamoDB and S3 with built-in durability. The service itself does not directly integrate with AWS Backup as there is no persistent session data to backup - each session starts fresh with applied policies and settings.

Disaster Recovery

WorkSpaces Secure Browser supports disaster recovery through its multi-region availability and stateless session architecture. Since sessions are ephemeral and configuration data is stored in AWS managed services, the service can be quickly restored in alternate regions. The service does not directly integrate with AWS Elastic Disaster Recovery as sessions are recreated on-demand rather than requiring instance-level recovery.

Recovery Objectives

WorkSpaces Secure Browser supports rapid recovery with minimal RPO and RTO through its ephemeral session model. Since no persistent data exists in browser sessions, RPO is effectively zero for session data. RTO is minimized through the service's ability to quickly provision new sessions and its multi-region availability, allowing rapid failover to alternate regions when needed.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/workspaces-web/latest/adminguide/request-service-quota.html

FAQ: https://aws.amazon.com/workspaces/secure-browser/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/workspaces-web/latest/adminguide/index.html

User Guide: https://docs.aws.amazon.com/workspaces-web/latest/adminguide/user-guide.html

API Reference: https://docs.aws.amazon.com/workspaces-web/latest/APIReference/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/workspaces/secure-browser/pricing/

Cost Optimization

WorkSpaces Secure Browser offers several cost optimization features including pay-as-you-go pricing with no minimum fees or long-term contracts, starting at $7 per month for standard.regular instances. The service includes up to 200 streaming hours per user monthly, with additional hours charged at competitive rates:

- Organizations can optimize costs by selecting appropriate instance types based on workload requirements - standard.regular for static content, standard.large for improved performance, and standard.xlarge for resource-intensive applications
- The ephemeral session model eliminates storage costs and reduces infrastructure overhead compared to traditional virtual desktop solutions

Savings Considerations

Primary cost savings come from eliminating VPN infrastructure, reducing virtual desktop licensing costs, and minimizing IT management overhead. The service eliminates the need for specialized client software, appliances, or dedicated infrastructure management:

- Organizations save on hardware procurement, maintenance, and upgrade cycles typically associated with traditional remote access solutions
- The fully managed service model reduces operational costs through automated patching, scaling, and capacity management
- Additional savings result from improved security posture reducing potential breach-related costs and simplified BYOD implementation without additional security infrastructure investments

Service Name

AWS Amplify

Service Overview

AWS Amplify is a comprehensive set of tools and services for building secure, scalable fullstack web and mobile applications. It provides frontend libraries, UI components, backend building tools, and hosting capabilities that enable developers to create cloud-connected applications with minimal configuration. Amplify offers a code-first developer experience where developers can define their backend using TypeScript and automatically provision AWS services like authentication, APIs, storage, and hosting. The platform supports multiple frameworks including React, Angular, Vue, Next.js, Flutter, React Native, iOS, and Android, making it versatile for various development needs.

Key Use Cases

- Fullstack web applications: Build complete web apps with authentication, real-time data, and hosting
- Mobile applications: Develop iOS, Android, Flutter, and React Native apps with cloud backend integration
- Real-time collaborative applications: Create messaging apps, project management tools, and chat applications with real-time data synchronization
- Content management systems: Build apps with file storage, user authentication, and content editing capabilities
- E-commerce platforms: Develop shopping apps with user accounts, product catalogs, and payment processing
- Social media applications: Create apps with user profiles, content sharing, and real-time interactions

Features

- **Authentication**: User sign-up, sign-in, and authorization using Amazon Cognito with support for SAML, OIDC, and social providers
- **Real-time Data APIs**: GraphQL and REST APIs powered by AWS AppSync with real-time subscriptions and offline synchronization
- **Storage**: File storage capabilities using Amazon S3 with different access levels and content management features
- **Hosting**: Fully managed hosting for static sites and server-side rendered apps with CI/CD and global CDN
- **DataStore**: Local data store with automatic cloud synchronization for offline-first applications

- **Functions**: Serverless functions using AWS Lambda for custom business logic and API extensions
- **Analytics**: User analytics and app usage tracking for understanding user behavior
- **Push Notifications**: Send targeted push notifications to mobile app users

Value Proposition

Customers should choose AWS Amplify for its comprehensive fullstack development capabilities that eliminate the complexity of configuring and managing multiple AWS services individually. Amplify provides a unified developer experience with type-safe, code-first backend definition using TypeScript, automatic provisioning of AWS resources, and seamless integration between frontend and backend:

- The platform offers built-in best practices for security, scalability, and performance while supporting multiple development frameworks and platforms
- Amplify's real-time capabilities, offline synchronization, and automatic scaling make it ideal for modern applications requiring responsive user experiences and global reach

Service Integration

AWS Amplify integrates deeply with core AWS services to provide comprehensive backend functionality. It uses Amazon Cognito for user authentication and authorization, AWS AppSync for GraphQL APIs and real-time subscriptions, Amazon DynamoDB for NoSQL database storage, and Amazon S3 for file storage:

- Amplify leverages AWS Lambda for serverless functions, Amazon CloudFront for global content delivery, and AWS CloudFormation for infrastructure provisioning
- The service is built on the AWS CDK, enabling seamless integration with any AWS service and supporting existing data sources including PostgreSQL and MongoDB
- This integration provides automatic scaling, security, and reliability across all backend components

Onboarding and Offboarding

Customers get started with AWS Amplify by installing the Amplify CLI and creating a new project using the quickstart guides available for different frameworks. The process involves configuring AWS credentials, defining the backend using TypeScript in the amplify/backend/ directory, and deploying resources using 'npx ampx sandbox' for development or 'npx ampx pipeline-deploy' for production:

- Amplify Studio provides a visual interface for non-technical users to manage content and configure backends

- The platform offers comprehensive documentation, tutorials, and sample applications for React, Angular, Vue, Next.js, Flutter, React Native, iOS, and Android development

Getting Started Guide: https://docs.amplify.aws/javascript/start/

Data Removal

Data removal and offboarding in AWS Amplify involves deleting the app from the Amplify console through App settings > General settings. This process removes all associated resources including hosting, authentication, APIs, and storage. For comprehensive cleanup, customers should delete CloudFormation stacks, remove S3 buckets, and clean up any custom resources created. The offboarding process is straightforward and takes less than 2 minutes to complete through the console interface.

Backup/Restore and Disaster Recovery

AWS Amplify provides backup and disaster recovery through its underlying AWS services. DynamoDB tables support point-in-time recovery and automated backups. S3 storage offers cross-region replication and versioning capabilities. Amplify applications can leverage AWS Backup for centralized backup management across resources. The CDK-based architecture allows configuration of deletion protection and backup policies for critical resources.

Backup Capabilities

Amplify supports backup capabilities through AWS CDK constructs that can enable point-in-time recovery for DynamoDB tables and configure AWS Backup for advanced backup options. Customers can modify resource configurations to enable deletion protection and automated backups on supported resources using underlying CDK construct properties.

Disaster Recovery

Amplify supports disaster recovery through its multi-region architecture and serverless foundation. Applications can be deployed across multiple AWS regions with automatic failover capabilities. The service leverages AWS AppSync for real-time data synchronization and Amazon CloudFront for global content delivery, providing resilience against regional failures.

Recovery Objectives

Amplify helps customers meet RTO and RPO objectives through its serverless architecture that automatically scales and recovers. Real-time data synchronization minimizes data loss, while global CDN distribution ensures rapid recovery. The multi-region deployment capabilities and automated backup features support various DR strategies from backup-and-restore to active-active configurations based on requirements.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/general/latest/gr/amplify.html

FAQ: https://aws.amazon.com/amplify/faqs/

Technical Requirements

Service Documentation: https://docs.amplify.aws/

User Guide: https://docs.amplify.aws/javascript/start/

API Reference: https://aws-amplify.github.io/amplify-js/api/

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/amplify/pricing/

Cost Optimization

AWS Amplify offers unique cost optimization through its serverless architecture that automatically scales resources based on demand, eliminating idle capacity costs. The service provides pay-per-use pricing for hosting, builds, and backend resources:

- Customers can optimize costs by leveraging the AWS Free Tier with up to $200 in credits for new accounts, implementing efficient caching strategies, and using appropriate storage classes
- The tiered pricing model for builds and hosting provides volume discounts, while the ability to pause sandbox environments reduces development costs
- Amplify's integration with AWS cost management tools enables detailed cost tracking and optimization recommendations

Savings Considerations

Main cost savings considerations include leveraging the generous free tier offering 12 months of free hosting and always-free development features. Customers save through reduced operational overhead as Amplify eliminates the need for server management, scaling, and infrastructure maintenance:

- The serverless architecture means paying only for actual usage rather than provisioned capacity
- Organizations can achieve significant cost reductions by consolidating multiple tools into Amplify's unified platform, reducing third-party service dependencies

- Development team productivity improvements through rapid prototyping and deployment capabilities translate to reduced development costs
- The automatic scaling prevents over-provisioning while ensuring performance during traffic spikes

Service Name

AWS Amplify Console

Service Overview

AWS Amplify Console is a unified management interface that provides a complete workflow for building, deploying, and hosting modern web applications and static sites. It combines Git-based CI/CD capabilities with global hosting on Amazon CloudFront's content delivery network. The console enables developers to deploy applications from code repositories (GitHub, GitLab, Bitbucket, AWS CodeCommit) with automatic builds on every commit, feature branch deployments, and pull request previews. It supports modern web frameworks including React, Angular, Vue, Next.js, Nuxt.js, and Gatsby, offering atomic deployments that eliminate maintenance windows and leveraging serverless technologies for automatic scaling to millions of users.

Key Use Cases

- Continuous deployment: Deploy web applications automatically from Git repositories with integrated CI/CD pipelines and feature branch deployments
- Static website hosting: Host single-page applications (SPAs) and static sites with global content distribution via Amazon CloudFront
- Full-stack application deployment: Deploy complete applications with both frontend and backend components using serverless infrastructure
- Development environment management: Create staging and production environments with pull request previews for testing changes
- Modern web framework deployment: Deploy applications built with React, Angular, Vue, Next.js, and other popular frameworks

Features

- **Git-based CI/CD**: Automated build and deployment pipeline triggered by Git commits with support for GitHub, GitLab, Bitbucket, and AWS CodeCommit
- **Feature Branch Deployments**: Automatic deployment of feature branches for testing and collaboration before merging to production
- **Pull Request Previews**: Generate preview environments for pull requests to test changes without affecting production
- **Global Hosting**: Deploy applications on Amazon CloudFront's global edge network for low-latency content delivery
- **Custom Domains & SSL**: Support for custom domains with automatic SSL certificate provisioning and wildcard subdomains
- **Atomic Deployments**: Zero-downtime deployments where sites are ready to view immediately after deployment completion

- **Password Protection**: Basic access authentication for protecting deployments during development and staging phases
- **Redirects & Rewrites**: Configure custom redirects, HTTP response codes, custom 404 pages, and proxy to external services
- **Real-time Monitoring**: Built-in hosting metrics monitoring with Amazon CloudWatch integration and custom alarms
- **AWS WAF Integration**: One-click web application firewall protection with managed rules for common exploits and custom security rules

Value Proposition

AWS Amplify Console offers a fully managed, serverless hosting solution that eliminates the operational overhead of managing web servers and infrastructure. Key differentiators include seamless integration with Git workflows for automatic CI/CD, atomic deployments that eliminate maintenance windows, and global content distribution through Amazon CloudFront:

- The service scales automatically to handle millions of users while maintaining pay-as-you-go pricing with no upfront costs
- Built-in security features like AWS WAF integration and SSL certificates reduce security management complexity
- The unified console experience provides developers with comprehensive tools for managing builds, hosting settings, deployed resources, and environment variables in a single interface, streamlining the entire application lifecycle from development to production

Service Integration

AWS Amplify Console integrates deeply with core AWS services to provide comprehensive full-stack application capabilities. Amazon CloudFront serves as the global content delivery network for hosting and distribution:

- AWS AppSync provides GraphQL APIs for real-time data synchronization
- Amazon Cognito handles user authentication and authorization with support for SAML and OIDC providers
- Amazon DynamoDB serves as the primary NoSQL database for scalable data storage
- AWS Lambda provides serverless compute for backend functions
- Amazon S3 stores static assets and build artifacts
- AWS CodeCommit supports Git repository hosting alongside external providers
- Amazon CloudWatch provides monitoring and logging capabilities
- AWS WAF offers web application firewall protection

- The service also leverages AWS CDK for infrastructure provisioning and CloudFormation templates for resource management and sharing between developers

Onboarding and Offboarding

Getting started with AWS Amplify Console is straightforward through the unified management interface. Customers can begin by logging into the Amplify console in their preferred AWS Region (available in 19 regions):

- The service offers quickstart templates and starter repositories that can be deployed directly from GitHub
- The typical onboarding flow involves connecting a Git repository (GitHub, GitLab, Bitbucket, or AWS CodeCommit), configuring build settings, and deploying the application
- Amplify automatically detects common frameworks and suggests appropriate build configurations
- For new projects, customers can use starter templates for popular frameworks like React, Next.js, Angular, or Vue
- The service provides comprehensive documentation, tutorials, and hands-on labs for each supported platform
- AWS offers up to $200 in free tier credits for new customers, making initial adoption cost-effective

Getting Started Guide: https://docs.amplify.aws/react/start/

Data Removal

Data removal involves deleting applications through the Amplify console, which removes associated build artifacts, deployment history, and configuration settings. Customers should manually delete connected Git repository webhooks and any custom domains before app deletion. All static assets stored in S3 and distributed through CloudFront are automatically cleaned up when applications are deleted. Environment variables and secrets are permanently removed. For complete offboarding, customers should also remove any backend resources provisioned through Amplify Studio, including DynamoDB tables, Lambda functions, and Cognito user pools, as these may incur ongoing charges if left active.

Backup/Restore and Disaster Recovery

AWS Amplify Console provides built-in resilience through its serverless architecture and global distribution via Amazon CloudFront. Build artifacts and deployment history are automatically stored and versioned, allowing rollback to previous deployments through the console. The service leverages AWS's multi-AZ infrastructure for high availability. Source code remains in the connected Git repositories, serving as the primary backup

mechanism. Static assets are automatically replicated across CloudFront edge locations globally, providing inherent disaster recovery capabilities through geographic distribution.

Backup Capabilities

Amplify Console maintains automatic versioning of build artifacts and deployment history, enabling rollback to previous application versions. The service does not directly integrate with AWS Backup as it primarily handles stateless web applications:

- However, any backend resources provisioned through Amplify (DynamoDB tables, S3 buckets) can be independently backed up using AWS Backup
- Git repositories serve as the primary source of truth for application code, providing version control and backup functionality

Disaster Recovery

While Amplify Console doesn't directly integrate with AWS Elastic Disaster Recovery, it provides inherent disaster recovery through its global, serverless architecture. Applications are automatically distributed across multiple CloudFront edge locations, ensuring availability even during regional outages:

- The atomic deployment model and Git-based source control enable rapid recovery and redeployment
- For comprehensive disaster recovery, customers should implement multi-region strategies for their backend services and ensure Git repositories are properly backed up

Recovery Objectives

Amplify Console supports aggressive Recovery Time Objectives (RTO) through its atomic deployment model and global distribution. Application rollbacks can be completed in minutes through the console interface. The serverless architecture eliminates single points of failure, typically achieving RTOs under 15 minutes for most incidents. Recovery Point Objectives (RPO) are primarily determined by Git commit frequency, as deployments are triggered by code commits. The global CloudFront distribution provides sub-second failover capabilities for content delivery.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/general/latest/gr/amplify.html

FAQ: https://aws.amazon.com/amplify/faqs/

Technical Requirements

Service Documentation: https://docs.amplify.aws/

User Guide: https://docs.amplify.aws/react/start/

API Reference: https://docs.aws.amazon.com/amplify/latest/APIReference/

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/amplify/pricing/

Cost Optimization

Amplify Console offers several cost optimization opportunities through its serverless, pay-as-you-go model. Build and deploy charges are based on actual usage rather than provisioned capacity:

- The service includes a generous free tier with ongoing free usage for build minutes and data transfer
- Cost optimization strategies include minimizing build artifact sizes to reduce storage costs, optimizing build times to reduce compute charges, using feature branch deployments judiciously to avoid unnecessary builds, and leveraging CloudFront's global caching to minimize data transfer costs
- The atomic deployment model eliminates the need for separate staging infrastructure, reducing overall hosting costs
- Environment-based pricing allows cost allocation across development, staging, and production environments

Savings Considerations

Primary cost savings come from eliminating traditional web server infrastructure and operational overhead. The serverless model means no ongoing server costs when applications aren't being accessed:

- Global CloudFront distribution reduces data transfer costs compared to single-region hosting
- Automatic scaling eliminates the need to overprovision for peak loads, with costs scaling directly with usage
- The integrated CI/CD pipeline reduces the need for separate build and deployment tools
- Free tier benefits include 1,000 build minutes per month and 15 GB of data served per month for hosting
- The service's atomic deployments eliminate the need for blue/green deployment infrastructure, further reducing costs
- Pay-per-use pricing for build minutes and data transfer ensures customers only pay for actual consumption rather than reserved capacity

Service Name

AWS AppFabric

Service Overview

AWS AppFabric is a fully managed service that quickly connects software-as-a-service (SaaS) applications across organizations to enhance security and increase productivity. It consists of two facets: AppFabric for security, which aggregates and normalizes audit logs from SaaS applications into a central repository using the Open Cybersecurity Schema Framework (OCSF), enabling IT and security teams to manage and secure applications using a standard schema; and AppFabric for productivity (preview), which leverages generative AI powered by Amazon Bedrock to enable employees to complete everyday tasks faster by automating task management and generating insights across their SaaS productivity applications. The service eliminates the need for custom code development and maintenance to connect applications.

Key Use Cases

- Use case 1: Security observability enhancement - Aggregating and normalizing audit logs from multiple SaaS applications to provide centralized security monitoring, standardized alerts, and improved incident response capabilities
- Use case 2: User access management - Quickly managing and auditing employee access across connected SaaS applications, enabling IT teams to see who has access to what applications through a single interface
- Use case 3: AI-powered productivity enhancement - Using generative AI to automatically facilitate tasks across applications, generate insights, prepare for meetings, and prioritize daily work activities

Features

- **Normalized Audit Logs**: Automatically normalizes SaaS application audit logs using the Open Cybersecurity Schema Framework (OCSF) into standardized formats (JSON or Apache Parquet)
- **Enriched Audit Logs**: Automatically enriches audit log data with user email addresses and additional context like IP addresses and user agent strings for comprehensive analysis
- **User Access Management**: Enables quick search and management of employee access across connected applications using corporate email addresses
- **Generative AI Assistant**: Powered by Amazon Bedrock, provides AI-driven insights and actions across applications with cross-app context
- **No-Code Integration**: Connects multiple SaaS applications without requiring custom code development or maintenance

- **Data Encryption**: Encrypts data at rest using AWS KMS keys and in transit using TLS 1.2 with AWS Signature V4

Value Proposition

AWS AppFabric provides unique value by eliminating the complexity and cost of building custom integrations between SaaS applications. It reduces operational costs by removing the need for custom code development and maintenance while automatically providing normalized audit logs in near real-time (every 2 minutes):

- The service offers enhanced security observability through standardized OCSF schema, enabling common security policies and standardized alerts across multiple applications
- AppFabric's generative AI capabilities powered by Amazon Bedrock provide cross-application insights and automation that competitors cannot match
- The service supports over 20 popular SaaS applications including Microsoft 365, Google Workspace, Slack, Zoom, and Salesforce, with compatibility for leading security tools like Splunk, Rapid7, and Netskope
- Organizations benefit from faster incident response times through enriched audit logs with user email addresses and reduced security blind spots across their SaaS portfolio

Service Integration

AWS AppFabric integrates deeply with several core AWS services to deliver its functionality. Primary integrations include Amazon S3 for storing aggregated audit log data, Amazon Kinesis Data Firehose for streaming normalized data to security tools and destinations, and AWS Key Management Service (KMS) for encryption of data at rest and customer-managed keys:

- The service leverages Amazon Bedrock to power its generative AI capabilities for productivity features
- AppFabric works with Amazon Security Lake to improve security posture by delivering OCSF-normalized data in Apache Parquet format
- Additional integrations include AWS CloudTrail for API and user activity logging, Amazon CloudWatch for metrics and monitoring, and AWS IAM for access control and service-linked roles
- The service also supports integration with Amazon Athena for querying audit log data and Amazon QuickSight for visualization and business intelligence dashboards

Onboarding and Offboarding

Customers get started with AppFabric by first creating an AWS account and navigating to the AppFabric console in supported regions (US East N. Virginia, Asia Pacific Tokyo, or Europe Ireland):

- The process begins with creating an app bundle, which serves as a container for app authorizations and ingestions
- During setup, customers configure encryption using either AWS-owned keys or customer-managed KMS keys
- Next, they authorize applications by providing necessary credentials like tenant IDs, client secrets, or personal access tokens for each SaaS application
- After authorization, customers set up audit log ingestions to specify destinations (Amazon S3 or Kinesis Data Firehose) for normalized data
- The user access tool can then be configured for employee access management
- AppFabric provides step-by-step guidance in the console and comprehensive documentation
- The service offers a 30-day free tier for the first two applications to facilitate proof-of-concept deployments

Getting Started Guide:
https://docs.aws.amazon.com/appfabric/latest/adminguide/getting-started-security.html

Data Removal

When offboarding from AppFabric, customers can remove their data by deleting app bundles, which permanently removes all associated metadata. Audit log data stored in customer-owned Amazon S3 buckets or delivered through Kinesis Data Firehose remains under customer control and must be deleted separately according to their data retention policies. AppFabric automatically retires issued grants on customer-managed KMS keys when app bundles are deleted. Intermediate S3 buckets used for temporary data processing are automatically deleted after 30 days using lifecycle policies. Customers should also revoke application authorizations and remove AppFabric access from their SaaS applications.

Backup/Restore and Disaster Recovery

AppFabric does not provide built-in backup and disaster recovery capabilities as it is a data aggregation and normalization service. The service relies on the underlying AWS infrastructure's resilience and availability. Customer audit log data is stored in Amazon S3 or streamed through Kinesis Data Firehose to customer-specified destinations, where standard AWS service backup and recovery features apply. AppFabric metadata and configurations are managed by AWS, and customers should document their app bundle configurations, authorizations, and ingestion settings for disaster recovery planning purposes.

Backup Capabilities

AppFabric does not have native backup capabilities as it functions as a data processing and aggregation service. The service does not store long-term customer data that would require backup:

- Audit log data flows through AppFabric to customer-specified destinations (Amazon S3 or Kinesis Data Firehose), where customers can implement their own backup strategies using AWS Backup or S3 versioning and replication features
- AppFabric configurations and authorizations should be documented and can be recreated as needed

Disaster Recovery

AppFabric does not directly support AWS Elastic Disaster Recovery as it is not a compute-intensive service requiring replication of instances or volumes. The service operates across multiple AWS regions (US East N:

- Virginia, Asia Pacific Tokyo, Europe Ireland) with AWS-managed high availability
- Customers should implement disaster recovery strategies for their audit log destinations in Amazon S3 or Kinesis Data Firehose using cross-region replication, backup policies, and documentation of AppFabric configurations for recreation in alternate regions if needed

Recovery Objectives

AppFabric helps customers meet RPO and RTO objectives by providing continuous audit log ingestion every two minutes, minimizing data loss in security monitoring scenarios. The service's near real-time data processing reduces the recovery point objective for security incident detection and response. However, since AppFabric is a data processing service without persistent storage, traditional RPO/RTO metrics are less applicable. Customers achieve their recovery objectives through the underlying AWS infrastructure's availability and by implementing appropriate backup and recovery strategies for their chosen data destinations.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/appfabric/latest/adminguide/limits.html

FAQ: https://aws.amazon.com/appfabric/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/appfabric/latest/adminguide/what-is-appfabric.html

User Guide: https://docs.aws.amazon.com/appfabric/latest/adminguide/getting-started-security.html

API Reference: https://docs.aws.amazon.com/appfabric/latest/api/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/appfabric/pricing/

Cost Optimization

AppFabric offers several cost optimization strategies specific to SaaS security management. The service eliminates the need for custom integration development and maintenance costs, providing significant operational savings:

- Customers benefit from a pay-as-you-go pricing model at $3 per user per month with volume discounts available for organizations with over 2,000 users
- The 30-day free tier for the first two applications allows for cost-effective proof-of-concept testing
- Organizations can optimize costs by carefully selecting which applications to connect based on security priorities
- Data storage costs can be managed by implementing appropriate S3 lifecycle policies and choosing between JSON and more storage-efficient Apache Parquet formats
- The user access feature pricing is per-search rather than per-user, allowing cost control based on actual usage patterns

Savings Considerations

The primary cost savings from AppFabric come from eliminating the need to build and maintain custom integrations between SaaS applications, which can cost hundreds of thousands of dollars in development and ongoing maintenance. Organizations save on security analyst time through normalized, enriched audit logs that accelerate incident response and reduce investigation time:

- The centralized user access management reduces IT overhead for access auditing and compliance reporting across multiple applications
- By providing standardized OCSF schema, AppFabric reduces the complexity and cost of security tool implementations and maintenance
- The generative AI productivity features can significantly reduce time spent on routine tasks like meeting preparation and task prioritization
- Volume pricing discounts and the ability to start with a subset of applications allow organizations to scale cost-effectively based on demonstrated value and ROI

Service Name

AWS Artifact

Service Overview

AWS Artifact provides on-demand downloads of AWS security and compliance documents, including reports on compliance with ISO standards, PCI Security Standards, and SOC reports. The service also offers downloads of certifications from accreditation bodies and documents for independent software vendors (ISVs) on AWS Marketplace. Additionally, AWS Artifact enables users to review, accept, and track the status of their agreements with AWS for individual accounts and multiple accounts within their organization, serving as a central resource for compliance-related information.

Key Use Cases

- Audit and compliance: Providing auditors and regulators with audit artifacts to demonstrate security and compliance of AWS infrastructure and services
- Vendor assessments: Eliminating the need to contact AWS Support during audits by providing self-service access to compliance reports and previous versions
- Organizational compliance management: Managing and accepting agreements on behalf of all member accounts in an AWS Organization through centralized controls

Features

- **On-demand compliance reports**: Download security and compliance documents including ISO, PCI, and SOC reports with unique watermarks
- **Agreement management**: Review, accept, and track status of AWS agreements for individual and organizational accounts
- **Third-party vendor reports**: Access compliance documents for independent software vendors selling products on AWS Marketplace
- **Historical report versions**: Access previous versions of compliance reports directly through the console for audit continuity
- **AWS Organizations integration**: Automatically accept security compliance reports on behalf of all member accounts in an organization

Value Proposition

AWS Artifact eliminates manual processes in compliance workflows by providing self-service access to security and compliance documentation, reducing dependency on AWS Support during audits. The service offers centralized agreement management across AWS Organizations, enabling scalable compliance operations:

- All documents and agreements are provided free of charge, making it a cost-effective solution for meeting regulatory requirements
- The service maintains audit trail integrity with watermarked documents and provides access to historical report versions, ensuring continuity during extended audit processes

Service Integration

AWS Artifact integrates closely with AWS Organizations to enable centralized agreement management and automatic acceptance of compliance reports across member accounts. It uses AWS Identity and Access Management (IAM) for fine-grained access control and service-linked roles:

- The service connects with AWS Marketplace to provide third-party vendor compliance documents
- Integration with AWS Audit Manager helps automate evidence collection for compliance frameworks
- AWS Artifact also works with AWS CloudTrail for API logging and audit trails

Onboarding and Offboarding

Getting started with AWS Artifact requires only an AWS account. Users access the AWS Artifact console to download reports, manage agreements, and configure notifications:

- The service provides immediate access to compliance documents without additional setup
- For AWS Organizations integration, administrators must enable trusted access and grant permissions for the service-linked role
- Prerequisites include setting up IAM permissions for users who need access to specific report types or agreement management functions

Getting Started Guide: https://docs.aws.amazon.com/artifact/latest/ug/getting-started.html

Data Removal

AWS Artifact operates as a document access service where users download compliance reports and manage agreements. Since documents are downloaded to user systems, data removal involves deleting locally stored files. For organizational agreements, administrators can terminate agreements through the console. The service maintains minimal customer data as it primarily provides access to AWS compliance documentation rather than storing customer-specific information.

Backup/Restore and Disaster Recovery

AWS Artifact does not provide traditional backup and disaster recovery capabilities as it serves as a document access portal for compliance reports and agreements. The service maintains high availability through AWS infrastructure, ensuring consistent access to compliance documentation. Report availability and agreement status are maintained by AWS across multiple regions for service reliability.

Backup Capabilities

AWS Artifact does not have traditional backup capabilities and does not integrate with AWS Backup, as it is primarily a document access service for compliance reports and agreements. The service focuses on providing reliable access to compliance documentation rather than data backup functionality.

Disaster Recovery

AWS Artifact does not support disaster recovery scenarios through AWS Elastic Disaster Recovery integration, as it serves as a compliance document access portal rather than a data processing service. Service continuity is maintained through AWS infrastructure resilience across regions.

Recovery Objectives

AWS Artifact supports compliance-related recovery objectives by providing consistent access to compliance documentation and agreements required during audit processes. The service maintains high availability to ensure auditors can access required documentation when needed, though it does not directly contribute to traditional RPO/RTO objectives for data recovery.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/general/latest/gr/aws-service-information.html

FAQ: https://aws.amazon.com/artifact/faq/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/artifact/latest/ug/what-is-aws-artifact.html

User Guide: https://docs.aws.amazon.com/artifact/latest/ug/what-is-aws-artifact.html

API Reference: https://docs.aws.amazon.com/artifact/latest/APIReference/

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/artifact/faq/

Cost Optimization

AWS Artifact provides exceptional cost optimization as all documents and agreements are offered free of charge to AWS customers. This eliminates traditional costs associated with compliance document procurement and agreement management:

- The service reduces operational costs by providing self-service access, eliminating the need for AWS Support interactions during audit processes
- Organizations can achieve economies of scale through centralized agreement management across multiple AWS accounts without additional fees

Savings Considerations

The primary cost savings from AWS Artifact come from eliminating manual compliance processes and reducing administrative overhead. Self-service access to compliance reports reduces time-to-audit completion and eliminates support request costs:

- Centralized agreement management through AWS Organizations reduces legal and administrative costs across multiple accounts
- The service's free tier model means organizations can access comprehensive compliance documentation without procurement budgets, making it particularly valuable for cost-conscious compliance programs

Service Name

AWS Audit Manager

Service Overview

AWS Audit Manager is a service that helps organizations continually audit their AWS usage to simplify how they assess risk and compliance with regulations and industry standards. Audit Manager automates evidence collection for policies, procedures, and activities (controls), making it easier to assess whether controls are operating effectively. The service provides prebuilt frameworks that structure and automate assessments for specific compliance standards or regulations, including a collection of controls with descriptions and testing procedures. When creating an assessment, Audit Manager automatically collects data from defined in-scope AWS accounts and transforms it into audit-friendly evidence for demonstrating compliance across security, change management, business continuity, and software licensing domains.

Key Use Cases

- Compliance Officers: Find non-compliant evidence across assessments and delegate remediation tasks to partner teams using evidence finder features
- Internal Risk Assessments: Support continuous auditing and compliance monitoring by automating evidence collection and maintaining audit readiness
- Multi-Account Evidence Management: Centrally manage and collect evidence from hybrid, multicloud, and multi-account AWS environments for comprehensive audit coverage

Features

- **Prebuilt Frameworks**: Gallery of frameworks supporting compliance standards like PCI DSS, HIPAA, SOC 2, GDPR, and AWS operational best practices
- **Automated Evidence Collection**: Continuous collection of evidence from AWS CloudTrail, AWS Config, AWS Security Hub, and API calls with configurable frequencies
- **Custom Controls and Frameworks**: Create custom controls from scratch or editable copies of existing controls to meet unique compliance requirements
- **Evidence Finder**: Search functionality to quickly find relevant evidence across all member accounts using filters and groupings
- **Delegation Workflow**: Assign control sets to subject matter experts for review, enabling cross-team collaboration in audit processes
- **Assessment Reports**: Generate audit-ready reports with evidence summaries and integrity validation through checksums

Value Proposition

AWS Audit Manager significantly reduces manual audit preparation effort by automating evidence collection and organizing it according to compliance requirements. The service saves thousands of hours typically spent manually collecting audit evidence, allowing teams to focus on risk remediation and audit planning:

- With prebuilt frameworks mapping AWS resources to control requirements for well-known standards, organizations can quickly establish audit readiness
- The delegation features enable efficient cross-team collaboration, while continuous evidence collection ensures ongoing compliance posture visibility
- Evidence integrity is maintained through secure, read-only storage with checksum validation, providing auditors with reliable, unaltered evidence for compliance assessments

Service Integration

AWS Audit Manager integrates deeply with multiple core AWS services for comprehensive evidence collection. AWS Security Hub provides automated security checks and compliance findings that Audit Manager captures as evidence snapshots:

- AWS Config supplies configuration compliance data and rule evaluations for resource assessments
- AWS CloudTrail delivers user activity logs that are continuously processed into audit evidence
- Additional integrations include AWS Organizations for multi-account assessments, AWS License Manager for software licensing compliance, AWS Control Tower for governance controls, and AWS Artifact for compliance documentation
- The service also connects with Amazon S3 for evidence storage and report generation, and supports integration with third-party GRC systems through APIs

Onboarding and Offboarding

Getting started requires completing account prerequisites including AWS account setup, creating administrative users, and granting necessary IAM permissions. Customers enable Audit Manager through the console, CLI, or API, with options for custom encryption keys and delegated administrator accounts:

- Recommended setup includes enabling AWS Organizations for multi-account assessments, AWS Security Hub for security findings, and AWS Config for compliance checks
- The onboarding process involves selecting from prebuilt frameworks or creating custom ones, specifying assessment scope across AWS accounts, designating audit owners, and initiating evidence collection

- First-time users should review Audit Manager concepts, complete the tutorial for creating assessments, and configure recommended integrations with supporting services

Getting Started Guide: https://docs.aws.amazon.com/audit-manager/latest/userguide/getting-started.html

Data Removal

Data removal occurs through multiple mechanisms: automatically when disabling the service, after a two-year retention period in managed storage repositories, or manually by deleting individual resources. Service metadata is encrypted at rest using AWS owned keys, while customer content uses either customer-managed or AWS owned keys. Evidence data is stored in Audit Manager's managed repository with read-only permissions, and assessment reports are saved to customer S3 buckets for additional control over data lifecycle management.

Backup/Restore and Disaster Recovery

AWS Audit Manager stores evidence in its own managed storage repository with built-in durability and availability. Evidence data is retained for up to two years with automatic deletion thereafter. Assessment reports are generated and stored in customer-specified S3 buckets, allowing customers to leverage S3's backup and versioning capabilities. The service does not provide native backup features but relies on AWS infrastructure resilience and customer-controlled S3 storage for report preservation and recovery scenarios.

Backup Capabilities

AWS Audit Manager does not have dedicated backup capabilities and does not integrate with AWS Backup service. The service maintains evidence in managed storage repositories with built-in durability, but customers must rely on S3 storage for assessment reports and implement their own backup strategies for critical compliance documentation. Evidence integrity is maintained through checksums and secure storage rather than traditional backup mechanisms.

Disaster Recovery

AWS Audit Manager does not integrate with AWS Elastic Disaster Recovery service. The service operates as a regional service ensuring evidence collection remains within regional boundaries:

- Disaster recovery relies on AWS infrastructure resilience and the distributed nature of evidence collection from multiple AWS services
- Customers must enable Audit Manager in each region separately and cannot aggregate evidence across regions, requiring region-specific disaster recovery planning

Recovery Objectives

AWS Audit Manager supports RPO objectives through continuous evidence collection from integrated services like CloudTrail, Config, and Security Hub, ensuring minimal data loss in compliance tracking. RTO objectives are supported by the service's regional architecture and integration with resilient AWS services. However, recovery depends on underlying AWS service availability and evidence collection resumption rather than specific RPO/RTO guarantees from Audit Manager itself.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/audit-manager/latest/userguide/service-quotas.html

FAQ: https://aws.amazon.com/audit-manager/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/audit-manager/latest/userguide/what-is.html

User Guide: https://docs.aws.amazon.com/audit-manager/latest/userguide/what-is.html

API Reference: https://docs.aws.amazon.com/audit-manager/latest/APIReference/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/audit-manager/pricing/

Cost Optimization

AWS Audit Manager offers a free tier providing 35,000 resource assessments per month for two calendar months for first-time customers. Cost optimization strategies include carefully scoping assessments to necessary accounts and resources to minimize resource assessment charges:

- Organizations can optimize costs by using common controls across multiple frameworks, leveraging shared custom frameworks across accounts and regions, and strategically scheduling evidence collection frequencies (daily, weekly, monthly) based on compliance requirements
- The evidence finder feature incurs additional CloudTrail Lake charges, so selective enablement based on search needs can control costs

Savings Considerations

Primary cost savings come from reduced manual audit preparation effort, potentially saving thousands of hours in evidence collection and organization. The automated evidence collection eliminates need for dedicated staff to manually gather compliance documentation:

- Using prebuilt frameworks reduces custom development costs compared to building compliance systems from scratch
- Multi-account evidence consolidation reduces operational overhead across large organizations
- However, customers should monitor resource assessment volumes as the primary cost driver, optimize assessment scope to essential resources, and leverage the free tier for initial evaluation and smaller workloads to maximize cost efficiency

Service Name

AWS B2B Data Interchange

Service Overview

AWS B2B Data Interchange is a fully managed service that automates the transformation and generation of Electronic Data Interchange (EDI) documents between JSON and XML formats. The service enables businesses to exchange transactional data with trading partners using standardized formats like X12, EDIFACT, or HL7v2. It provides generative AI-assisted, bi-directional EDI transformation, event-driven data integration, trading partner management, and built-in HIPAA compliance. The service monitors Amazon S3 locations for new EDI documents and automatically transforms them, reducing the time, complexity, and cost associated with managing EDI data while enabling integration with AWS analytics, AI, and machine learning services.

Key Use Cases

- Healthcare data exchange: Transform HL7v2 and X12 EDI documents for HIPAA-compliant healthcare transactions between providers, payers, and systems
- Supply chain and logistics: Automate processing of purchase orders, invoices, advance ship notices, and inventory management documents with trading partners
- Financial services: Process payment transactions, remittance advice, and financial reporting documents in standardized EDI formats

Features

- **Generative AI-assisted EDI mapping**: Uses Amazon Bedrock and Anthropic Claude models to automatically generate mapping code for bi-directional EDI transformations
- **Event-driven processing**: Monitors Amazon S3 locations and generates EventBridge events for downstream integrations and workflow automation
- **Automatic acknowledgment generation**: Generates 997, 999, and TA1 functional acknowledgments for inbound X12 EDI transactions automatically
- **Trading partner management**: Manages partnerships with profile information, trading capabilities, and customizable EDI header configurations
- **Bi-directional transformation**: Converts inbound EDI to JSON/XML and generates outbound EDI documents from JSON/XML data
- **HIPAA compliance**: Built-in HIPAA eligibility for healthcare organizations handling protected health information

Value Proposition

AWS B2B Data Interchange eliminates the high licensing fees, operational complexity, and lengthy onboarding processes associated with traditional EDI solutions. The service provides pay-as-you-go pricing with no upfront costs or infrastructure to manage:

- Key advantages include generative AI-assisted mapping that reduces the time and technical expertise required for EDI implementations, automatic acknowledgment generation, event-driven processing for seamless integration with downstream systems, and built-in HIPAA compliance
- The fully managed nature removes operational burdens while enabling businesses to focus on gaining insights from their transactional data using AWS analytics and machine learning services

Service Integration

AWS B2B Data Interchange integrates closely with Amazon S3 for document storage and monitoring, Amazon EventBridge for event-driven workflows, and AWS Transfer Family for secure document exchange via SFTP and AS2 protocols. The service leverages Amazon Bedrock and Anthropic Claude models for generative AI-assisted mapping capabilities:

- Additional integrations include Amazon CloudWatch for logging and monitoring, Amazon SNS for notifications, AWS Lambda for custom processing, AWS Step Functions for workflow orchestration, and AWS Glue for ETL operations
- The service also works with AWS security services like Amazon GuardDuty for malware scanning and supports various Amazon S3 encryption options

Onboarding and Offboarding

Getting started requires creating four main components: a profile with organization contact information, a transformer with EDI transaction details and optional custom mapping code, a trading capability specifying EDI direction and input/output directories, and a partnership with trading partner details and contextual information. Customers can write mapping code manually or use generative AI-assisted capabilities through Amazon Bedrock:

- The process involves uploading sample EDI and data files to Amazon S3, configuring these components through the AWS Management Console, and then dropping documents into specified input directories for automatic transformation
- The service provides comprehensive documentation, tutorials, and a self-paced workshop to guide implementation

Getting Started Guide: https://docs.aws.amazon.com/b2bi/latest/userguide/getting-started.html

Data Removal

EDI documents, acknowledgments, and transformed data files are stored in customer-owned Amazon S3 buckets, providing full control over data retention and deletion. Customers can delete data directly from their S3 buckets following standard S3 deletion procedures. Service configurations including profiles, transformers, trading capabilities, and partnerships can be removed through the AWS Management Console or APIs. The service does not store customer data for training purposes, ensuring complete data privacy during offboarding.

Backup/Restore and Disaster Recovery

AWS B2B Data Interchange inherits backup and disaster recovery capabilities through its underlying AWS infrastructure and Amazon S3 storage. Customer data stored in S3 benefits from S3's built-in durability and availability features. The service configuration and metadata can be recreated using Infrastructure as Code approaches like AWS CloudFormation. However, the service does not provide native backup or disaster recovery features beyond those offered by the underlying AWS services it depends upon.

Backup Capabilities

The service does not have built-in backup capabilities and does not integrate with AWS Backup. Data protection relies on Amazon S3's inherent durability and versioning capabilities where documents are stored:

- Customers must implement their own backup strategies for S3 data if additional protection is required
- Service configurations can be backed up through Infrastructure as Code templates or by exporting configuration settings manually

Disaster Recovery

AWS B2B Data Interchange does not have native disaster recovery capabilities and does not integrate with AWS Elastic Disaster Recovery. Disaster recovery relies on the multi-Availability Zone architecture of the underlying AWS services. Customers must implement cross-region replication strategies for their S3 data and recreate service configurations in alternate regions if regional disaster recovery is required.

Recovery Objectives

The service can support customer RPO and RTO objectives through careful architecture design using Amazon S3 cross-region replication for data protection and AWS CloudFormation templates for rapid service configuration deployment. Recovery objectives depend on the implementation of supporting services like S3 replication frequency and the time required to deploy configurations in alternate regions. The service's event-driven architecture can help minimize processing delays during recovery scenarios.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/b2bi/latest/userguide/b2bi-quotas.html

FAQ: https://aws.amazon.com/b2b-data-interchange/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/b2bi/latest/userguide/index.html

User Guide: https://docs.aws.amazon.com/b2bi/latest/userguide/what-is-b2bi.html

API Reference: https://docs.aws.amazon.com/b2bi/latest/APIReference/index.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/b2b-data-interchange/pricing/

Cost Optimization

AWS B2B Data Interchange offers several cost optimization opportunities including pay-as-you-go pricing with no upfront costs or infrastructure to manage. Customers pay only for active partnerships and transformation steps used, avoiding traditional EDI solution licensing fees:

- The generative AI mapping feature incurs Amazon Bedrock charges only during mapping generation, not during runtime transformations
- Organizations can optimize costs by carefully managing the number of active partnerships, using default EDI transformations when custom mapping isn't required, and leveraging Amazon S3 storage classes for long-term document retention
- The service eliminates operational overhead costs associated with traditional EDI infrastructure maintenance and scaling

Savings Considerations

Primary cost savings come from eliminating traditional EDI solution licensing fees, infrastructure maintenance costs, and operational overhead. The service's fully managed nature reduces staffing requirements for EDI operations and system administration:

- Generative AI-assisted mapping significantly reduces development time and costs compared to manual mapping creation
- Event-driven automation reduces manual processing costs and errors

- Organizations save on hardware, software licensing, maintenance contracts, and specialized EDI expertise typically required for on-premises solutions
- The pay-as-you-use model eliminates costs for unused capacity and allows scaling based on actual business needs, providing significant savings for organizations with variable EDI volumes

Service Name

AWS Backup

Service Overview

AWS Backup is a fully managed service that centralizes and automates data protection across AWS services, in the cloud, and on-premises. It allows users to configure backup policies and monitor backup activity for AWS resources in one place, removing the need for custom scripts and manual processes. The service consolidates backup tasks that were previously performed service-by-service, enabling consistent and compliant backup strategies across organizations. AWS Backup supports policy-based backup management, automated scheduling, and cross-Region/cross-account backup capabilities for comprehensive data protection.

Key Use Cases

- Centralized backup management: Consolidate backup policies and monitoring across multiple AWS services from a single console
- Regulatory compliance: Meet business and regulatory requirements with automated backup policies, audit reporting, and WORM capabilities
- Disaster recovery: Implement cross-Region backup replication and automated recovery testing to ensure business continuity
- Ransomware protection: Use logically air-gapped vaults, vault lock, and malware scanning to protect against cyber threats
- Multi-account governance: Apply organization-wide backup policies across AWS Organizations with centralized monitoring and reporting

Features

- **Centralized Backup Management**: Single console to manage backup policies and monitor activity across all AWS services
- **Policy-Based Backup**: Create backup plans with automated scheduling, retention, and lifecycle management
- **Tag-Based Backup Policies**: Apply backup plans to resources using AWS tags for scalable backup strategy implementation
- **Cross-Region and Cross-Account Backup**: Automatically replicate backups across AWS Regions and accounts for enhanced security
- **AWS Backup Audit Manager**: Monitor compliance posture and generate reports for governance and regulatory requirements
- **Backup Vault Lock (WORM)**: Write-once-read-many configuration prevents deletion or modification of backups

- **Incremental Backups**: Efficient storage with full initial backups followed by incremental changes only
- **Backup Search and Item-Level Restore**: Search and restore specific files or objects from EBS snapshots and S3 backups
- **Malware Protection**: Integration with Amazon GuardDuty for automated malware scanning of recovery points
- **Restore Testing**: Automated validation of backup recoverability to meet RTO/RPO targets

Value Proposition

AWS Backup provides the only fully managed, centralized backup solution across AWS services, eliminating the complexity of managing multiple backup tools and custom scripts. It offers enterprise-grade features like WORM compliance, cross-Region replication, and automated restore testing that would be costly and complex to build independently:

- The service integrates seamlessly with AWS Organizations for multi-account governance and provides comprehensive audit capabilities for regulatory compliance
- With pay-as-you-go pricing, no upfront costs, and automatic lifecycle management to reduce storage costs, AWS Backup delivers significant operational efficiency while ensuring data protection meets business and compliance requirements

Service Integration

AWS Backup integrates with over 20 AWS services including Amazon EC2, Amazon S3, Amazon EBS, Amazon RDS, Amazon Aurora, Amazon DynamoDB, Amazon EFS, Amazon FSx, Amazon DocumentDB, Amazon Neptune, Amazon Redshift, Amazon Timestream, and VMware workloads. It also integrates with AWS Organizations for multi-account management, AWS IAM for access control, AWS KMS for encryption, Amazon CloudWatch for monitoring, AWS CloudTrail for audit logging, and Amazon GuardDuty for malware protection. The service works with AWS Cost Explorer for cost management and AWS Config for compliance tracking.

Onboarding and Offboarding

Getting started with AWS Backup requires an AWS account and at least one supported AWS resource. Users must first opt-in AWS services through the AWS Management Console to enable AWS Backup management:

- The process involves creating backup plans that define schedules and retention policies, then assigning resources using tags or explicit resource selection
- AWS Backup provides default backup plans and policies to accelerate implementation

- Prerequisites include configuring appropriate IAM roles and policies, enabling service opt-in for specific resource types like Aurora and DocumentDB, and setting up backup vaults for organizing and securing backup data

Getting Started Guide: https://docs.aws.amazon.com/aws-backup/latest/devguide/getting-started.html

Data Removal

AWS Backup provides multiple options for data removal including automated lifecycle policies to delete expired backups and manual deletion of recovery points through the console or API. Users can permanently delete backup data or stop future backups while retaining existing data. For secure offboarding, Backup Vault Lock in Compliance mode prevents deletion during retention periods. Organizations can remove AWS Backup configurations while preserving data in underlying AWS services, ensuring complete data lifecycle management.

Backup/Restore and Disaster Recovery

AWS Backup provides comprehensive backup and disaster recovery capabilities including automated backup scheduling, cross-Region replication, point-in-time recovery for supported services, and restore testing validation. The service supports both continuous backups with sub-second RPO and periodic backups based on defined schedules. Recovery options include full resource restoration, item-level recovery, and cross-account recovery scenarios for comprehensive disaster recovery strategies.

Backup Capabilities

AWS Backup itself is a backup service that provides centralized backup management for other AWS resources. It works seamlessly with AWS Backup (itself) by providing the core backup infrastructure, policies, and management capabilities. The service offers incremental backups, encrypted storage, cross-Region replication, and automated lifecycle management to optimize backup storage costs and ensure data protection across all supported AWS services.

Disaster Recovery

AWS Backup supports disaster recovery through cross-Region backup replication, automated restore testing, and integration with AWS Elastic Disaster Recovery for complementary capabilities. While AWS Backup focuses on backup-based recovery (hours RTO/RPO), it complements AWS Elastic Disaster Recovery which provides continuous replication with sub-minute RTO/RPO. Together, these services provide comprehensive disaster recovery options for different recovery time objectives.

Recovery Objectives

AWS Backup helps customers meet RPO objectives through configurable backup frequencies from continuous (sub-second RPO) to periodic schedules. RTO objectives are supported through restore testing that validates recovery times, cross-Region backup copies for faster regional recovery, and item-level restore capabilities that reduce recovery scope. Automated restore testing ensures backup recoverability and validates that recovery processes meet defined RTO targets.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/aws-backup/latest/devguide/aws-backup-limits.html

FAQ: https://aws.amazon.com/backup/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/aws-backup/latest/devguide/whatisbackup.html

User Guide: https://docs.aws.amazon.com/aws-backup/latest/devguide/getting-started.html

API Reference: https://docs.aws.amazon.com/aws-backup/latest/devguide/api-reference.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/backup/pricing/

Cost Optimization

AWS Backup offers several unique cost optimization features including automated lifecycle management that transitions backups to lower-cost cold storage, backup tiering for Amazon S3 that reduces costs by up to 30% for long-term storage, and incremental backup technology that only stores changed data. The service provides backup deduplication, automated deletion of expired backups, and cost allocation tags for detailed cost tracking. Organizations can optimize costs through policy-based retention management and cross-Region backup strategies that balance cost with recovery requirements.

Savings Considerations

Primary cost savings from AWS Backup include elimination of custom backup solution development and maintenance costs, reduced storage expenses through incremental backups and lifecycle policies, and operational efficiency gains from centralized management. The service offers significant savings through automated backup scheduling that reduces manual labor, consolidated backup infrastructure that eliminates per-service backup tools, and pay-as-you-go pricing with no upfront costs. Additional savings come from backup tiering options, cross-Region cost optimization, and automated compliance reporting that reduces audit preparation time and costs.

Service Name

AWS Batch

Service Overview

AWS Batch is a fully managed batch computing service that helps users run batch computing workloads on the AWS Cloud. It eliminates the undifferentiated heavy lifting of configuring and managing required infrastructure, similar to traditional batch computing software. AWS Batch automatically provisions compute resources and optimizes workload distribution based on the quantity and scale of submitted workloads. The service runs containerized batch ML, simulation, and analytics workloads across various AWS compute offerings including Amazon ECS, Amazon EKS, AWS Fargate, and Spot or On-Demand Instances. AWS Batch provides queuing, scheduling, and priority management capabilities while delivering operational capabilities to optimize workloads for throughput, speed, resource efficiency, and cost.

Key Use Cases

- High Performance Computing (HPC): Running compute-intensive workloads for scientific research, engineering simulations, and complex calculations at scale
- Machine Learning and AI: Training ML models, hyperparameter tuning, and running inference workloads with automatic resource scaling and SageMaker integration
- Financial Services: Post-trade analytics for risk analysis, high-frequency trading simulations, and fraud surveillance using machine learning algorithms
- Life Sciences: Drug screening for molecular libraries, DNA sequencing analysis, and genomics research with large-scale computational requirements
- Digital Media: Video rendering, transcoding workflows, and media supply chain processing for content creation and distribution
- Scientific Research: Monte Carlo simulations, climate modeling, and data processing for research institutions and laboratories

Features

- **Fully Managed Service**: Automatically provisions, scales, and terminates compute resources based on job requirements without manual infrastructure management
- **Multi-Compute Environment Support**: Supports Amazon EC2, AWS Fargate, Amazon EKS, and Spot Instances for flexible compute resource allocation
- **Job Queuing and Scheduling**: Priority-based job scheduling with dependency management and automatic retry capabilities for failed jobs
- **Container Orchestration**: Runs containerized workloads using Docker containers with support for multi-container and multi-node parallel jobs

- **Resource-Aware Scheduling**: Manages consumable resources like license tokens and bandwidth limits to optimize job scheduling and resource utilization
- **SageMaker Integration**: Provides queuing and scheduling capabilities for Amazon SageMaker Training jobs with fair-share scheduling policies
- **Auto Scaling**: Dynamically scales compute resources up and down based on job queue demand and workload requirements
- **Cost Optimization**: Supports EC2 Spot Instances for up to 90% cost savings and integrates with AWS cost management tools

Value Proposition

AWS Batch provides significant advantages over traditional batch computing solutions by eliminating infrastructure management overhead and reducing operational complexity. Customers benefit from automatic scaling that adapts to workload demands without manual intervention, resulting in improved resource utilization and cost efficiency:

- The service integrates seamlessly with existing AWS services and supports multiple compute environments, allowing organizations to optimize performance and costs
- AWS Batch's managed approach reduces the need for specialized infrastructure teams, enabling data scientists and researchers to focus on their core work rather than infrastructure management
- The service also provides built-in fault tolerance, automatic retries, and comprehensive monitoring capabilities that enhance reliability and operational visibility

Service Integration

AWS Batch integrates deeply with core AWS services to provide comprehensive batch computing capabilities. It leverages Amazon EC2 for compute instances, AWS Fargate for serverless containers, and Amazon EKS for Kubernetes-based workloads:

- The service integrates with Amazon ECS for container orchestration and management
- For machine learning workloads, AWS Batch provides native integration with Amazon SageMaker for training job queuing and scheduling
- Storage integration includes Amazon S3 for data input/output, Amazon EFS for shared file systems, and Amazon EBS for persistent storage
- Monitoring and logging capabilities are provided through Amazon CloudWatch for metrics and logs, AWS CloudTrail for auditing, and AWS Trusted Advisor for cost optimization recommendations
- The service also integrates with AWS IAM for access control and AWS VPC for network isolation

Onboarding and Offboarding

Getting started with AWS Batch involves creating core components in sequence: compute environments, job queues, job definitions, and jobs. Customers first set up IAM roles and policies, then create a compute environment specifying the compute resources (EC2, Fargate, or EKS) and configuration:

- Next, they create job queues that reference the compute environment and define scheduling priorities
- Job definitions specify container properties, resource requirements, and execution parameters
- Finally, jobs are submitted to queues for execution
- AWS provides multiple tutorials including wizard-based setup for beginners, CLI-based tutorials for advanced users, and specific guidance for different compute environments
- The onboarding process includes setting up VPC networking, security groups, and IAM permissions
- AWS offers comprehensive documentation, tutorials, and best practices guides to help customers implement batch workloads effectively

Getting Started Guide:
https://docs.aws.amazon.com/batch/latest/userguide/Batch_GetStarted.html

Data Removal

AWS Batch offboarding involves deleting resources in reverse order of creation and removing associated data. Customers must first disable and delete job queues, then disable and delete compute environments. Job definitions and completed job records can be deregistered and removed. Any data stored in Amazon S3 buckets, EBS volumes, or EFS file systems used by batch jobs must be manually deleted. Container images stored in Amazon ECR should be removed if no longer needed. CloudWatch logs and metrics are automatically retained according to configured retention policies but can be manually deleted. IAM roles and policies created for AWS Batch can be removed after confirming no other services depend on them.

Backup/Restore and Disaster Recovery

AWS Batch relies on underlying AWS services for backup and disaster recovery capabilities. Job definitions and compute environment configurations are stored in AWS Batch service metadata and automatically replicated across availability zones. Input and output data stored in Amazon S3 can leverage S3 versioning, cross-region replication, and backup policies. EBS volumes used by compute instances can be backed up using EBS snapshots and AWS Backup service. Container images in Amazon ECR support cross-region replication for disaster recovery. AWS Batch workloads can be recreated in different regions using Infrastructure as Code tools like CloudFormation.

Backup Capabilities

AWS Batch does not provide native backup capabilities but integrates with AWS Backup for underlying resources. The service configuration including job definitions and compute environments is automatically maintained by AWS with high availability:

- Customers can use AWS Backup to protect EBS volumes, EFS file systems, and other storage resources used by batch jobs
- Container images can be backed up through Amazon ECR replication features
- Job execution logs in CloudWatch can be exported to S3 for long-term retention and backup

Disaster Recovery

AWS Batch supports disaster recovery through multi-region deployment capabilities and integration with AWS disaster recovery services. The service can be deployed across multiple AWS regions with job definitions and compute environments recreated using Infrastructure as Code:

- AWS Batch integrates with AWS Elastic Disaster Recovery for comprehensive disaster recovery planning
- Workloads can be designed to automatically failover to secondary regions using AWS services like Route 53 for DNS failover and Application Recovery Controller for automated recovery orchestration

Recovery Objectives

AWS Batch helps customers meet RPO and RTO objectives through automated scaling, multi-region deployment, and integration with AWS backup services. Recovery Point Objectives can be achieved by implementing regular backups of input data using S3 versioning and cross-region replication, plus EBS snapshots for persistent storage. Recovery Time Objectives are supported through rapid resource provisioning, pre-configured compute environments in multiple regions, and automated job restart capabilities. The service's ability to quickly scale compute resources and recreate batch processing environments helps minimize recovery time during disaster scenarios.

Service Constraints

Service Quotas:
https://docs.aws.amazon.com/batch/latest/userguide/service_limits.html

FAQ: https://aws.amazon.com/batch/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/batch/latest/userguide/what-is-batch.html

User Guide: https://docs.aws.amazon.com/batch/latest/userguide/

API Reference: https://docs.aws.amazon.com/batch/latest/APIReference/

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/batch/pricing/

Cost Optimization

AWS Batch offers unique cost optimization through automatic integration with EC2 Spot Instances for up to 90% cost savings compared to On-Demand pricing. The service automatically manages Spot Instance bidding, replacement, and interruption handling:

- Customers can leverage mixed instance types and allocation strategies to optimize cost and performance
- AWS Batch supports Savings Plans and Reserved Instances for predictable workloads
- Dynamic scaling ensures customers only pay for compute resources when jobs are running, eliminating idle capacity costs
- The service integrates with AWS Cost Explorer and AWS Budgets for cost monitoring and alerts
- Fair-share scheduling and resource-aware scheduling help optimize resource utilization and prevent over-provisioning

Savings Considerations

Primary cost savings come from eliminating the need to manage and maintain dedicated batch processing infrastructure, reducing operational overhead and staffing requirements. Automatic scaling prevents over-provisioning of resources and eliminates idle compute costs during low-demand periods:

- Spot Instance integration provides substantial cost reductions for fault-tolerant workloads with up to 90% savings over On-Demand pricing
- The service's efficient job scheduling and resource allocation minimize wasted compute cycles and improve overall resource utilization
- Multi-tenancy and shared infrastructure reduce per-job costs compared to dedicated solutions
- Integration with AWS cost management tools enables continuous optimization and cost control through monitoring, budgets, and recommendations

Service Name

AWS Budgets

Service Overview

AWS Budgets is a cloud financial management service that enables users to set custom budgets for AWS costs and usage, track Reserved Instances (RI) and Savings Plans utilization and coverage metrics, and receive notifications when spending approaches or exceeds defined thresholds. The service updates budget information up to three times a day and supports various budget types including cost, usage, RI utilization, RI coverage, Savings Plans utilization, and Savings Plans coverage budgets. Users can configure automated actions when budget thresholds are met, such as applying IAM policies or stopping EC2 instances, providing proactive cost control and governance.

Key Use Cases

- Cost Control and Monitoring: Set spending limits for AWS services and receive alerts when costs approach or exceed defined thresholds to prevent unexpected charges
- Usage Tracking and Governance: Establish usage limits for specific services and get notified when usage approaches set thresholds to stay within service limits
- Reserved Instances and Savings Plans Optimization: Monitor RI and Savings Plans utilization and coverage to identify unused or under-utilized commitments and maximize cost savings

Features

- **Cost Budgets**: Set spending limits for services and receive alerts when costs approach or exceed defined thresholds
- **Usage Budgets**: Establish usage limits for one or more services and get notified when usage approaches or exceeds set thresholds
- **RI Utilization Budgets**: Define utilization thresholds for Reserved Instances and receive alerts when usage falls below defined levels
- **RI Coverage Budgets**: Set coverage thresholds and get alerted when percentage of instance hours covered by RIs falls below defined levels
- **Savings Plans Budgets**: Establish utilization and coverage thresholds for Savings Plans and receive notifications when metrics drop below set levels
- **Budget Actions**: Configure automated actions when budget thresholds are met, such as applying IAM policies or stopping EC2 instances
- **Multiple Budget Methods**: Support for Fixed, Planned, and Auto-adjusting budget methods to accommodate different planning approaches

- **Custom Period Budgets**: Create budgets that align with fiscal years, project durations, or grant periods with custom time ranges
- **Forecasting Alerts**: Receive notifications based on forecasted spending before actual costs are incurred
- **Multi-dimensional Filtering**: Apply filters based on services, regions, accounts, tags, and other dimensions for granular budget tracking

Value Proposition

AWS Budgets provides comprehensive cost control and financial governance capabilities that are deeply integrated with AWS billing and cost management infrastructure. Unlike third-party solutions, AWS Budgets offers native integration with all AWS services, real-time cost and usage data updated up to three times daily, and automated actions that can directly control AWS resources:

- The service provides flexible budgeting methods including fixed, planned, and auto-adjusting approaches, supports forecasting alerts to prevent cost overruns before they occur, and enables granular tracking through multiple dimensions
- With no additional cost for basic budgeting and notifications, AWS Budgets delivers enterprise-grade financial management capabilities while maintaining simplicity for smaller organizations

Service Integration

AWS Budgets integrates seamlessly with core AWS services including AWS Cost Explorer for cost analysis and historical data, Amazon SNS for notification delivery, AWS IAM for access control and automated policy actions, AWS Organizations for consolidated billing and multi-account budget management, Amazon CloudWatch for monitoring and alerting, AWS Cost and Usage Reports for detailed billing data, and AWS Cost Anomaly Detection for identifying unusual spending patterns. The service also works with Reserved Instances and Savings Plans for commitment tracking, AWS Service Catalog for portfolio budget management, and supports automated actions on EC2 instances and RDS resources. This deep integration enables comprehensive financial governance across the entire AWS ecosystem.

Onboarding and Offboarding

Customers can get started with AWS Budgets by accessing the AWS Billing and Cost Management console and navigating to the Budgets page. The service offers simplified templates for common use cases like zero spend budgets, monthly cost budgets, and reservation utilization budgets, as well as advanced customization options for complex scenarios:

- Users can create budgets using either simplified templates or advanced customization methods, configure notifications through email addresses and SNS topics, and set up automated actions
- The service requires appropriate IAM permissions for viewing billing information, creating CloudWatch alarms, and managing SNS notifications
- Best practices recommend starting with basic cost budgets before implementing more complex usage and reservation budgets

Getting Started Guide: https://docs.aws.amazon.com/cost-management/latest/userguide/budgets-managing-costs.html

Data Removal

AWS Budgets data removal involves deleting individual budgets through the console or API, which removes all associated budget history, notifications, and actions. When customers delete budgets, the historical performance data is permanently removed and cannot be recovered. For complete offboarding, customers should delete all budgets, remove associated SNS topics and IAM policies if no longer needed, and ensure budget actions are disabled to prevent any automated resource management. Billing and usage data used by budgets remains available through other AWS billing services.

Backup/Restore and Disaster Recovery

AWS Budgets does not provide traditional backup and restore capabilities as it is a configuration-based service that tracks financial data rather than storing customer application data. Budget configurations and historical performance data are maintained by AWS across multiple availability zones for high availability. Customers can export budget history data in CSV format for external backup purposes and can recreate budget configurations using Infrastructure as Code tools like CloudFormation or programmatically through the AWS API.

Backup Capabilities

AWS Budgets does not have dedicated backup capabilities or integration with AWS Backup service, as it primarily manages configuration data and financial tracking information. The service maintains budget configurations and historical data with high availability across AWS infrastructure. Customers can download budget performance history as CSV files for external archival and can use CloudFormation templates or API calls to backup and restore budget configurations programmatically.

Disaster Recovery

AWS Budgets does not integrate with AWS Elastic Disaster Recovery as it is not a data-intensive service requiring disaster recovery replication. The service is inherently resilient as budget configurations and notifications are managed across multiple AWS availability zones. In case of regional issues, budget functionality would be restored automatically as

part of AWS service recovery, and customers can recreate budget configurations using saved CloudFormation templates or API scripts.

Recovery Objectives

AWS Budgets supports Recovery Point Objective (RPO) through downloadable CSV exports of budget history data, allowing customers to maintain external copies of budget performance information. For Recovery Time Objective (RTO), budget configurations can be quickly recreated using Infrastructure as Code templates or API automation. Since AWS Budgets is a configuration service rather than a data storage service, recovery focuses on recreating budget settings and thresholds rather than data restoration, enabling rapid restoration of financial governance controls.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/cost-management/latest/userguide/management-limits.html

FAQ: https://aws.amazon.com/aws-cost-management/aws-budgets/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/cost-management/latest/userguide/budgets-managing-costs.html

User Guide: https://docs.aws.amazon.com/cost-management/latest/userguide/budgets-managing-costs.html

API Reference: https://docs.aws.amazon.com/aws-cost-management/latest/APIReference/API_Operations_AWS_Budgets.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/aws-cost-management/aws-budgets/pricing/

Cost Optimization

AWS Budgets offers unique cost optimization through proactive financial governance rather than reactive cost cutting. The service enables automated responses to spending patterns through budget actions that can apply restrictive IAM policies or stop resources when thresholds are exceeded, preventing cost overruns before they occur:

- Customers can optimize Reserved Instance and Savings Plans investments by monitoring utilization and coverage metrics, identifying underutilized commitments for reallocation or modification

- The forecasting capabilities allow organizations to adjust spending patterns before reaching budget limits, and the granular filtering options enable identification of specific cost drivers across services, regions, and teams for targeted optimization efforts

Savings Considerations

AWS Budgets provides significant cost savings through early detection of spending anomalies and automated cost controls that prevent budget overruns. The service itself incurs minimal costs with free basic budgeting and notifications, and only charges $0.10 daily for action-enabled budgets beyond the first two free ones per month:

- Primary savings come from preventing unexpected charges through proactive alerts and automated resource management, optimizing Reserved Instance and Savings Plans utilization to maximize discount benefits, and enabling organizations to identify and eliminate wasteful spending patterns through detailed budget analysis
- The forecasting capabilities allow businesses to make informed purchasing decisions and avoid emergency procurement at higher on-demand rates

Service Name

AWS Certificate Manager

Service Overview

AWS Certificate Manager (ACM) is a service that simplifies the process of creating, managing, and renewing SSL/TLS X.509 certificates and their corresponding private keys. ACM manages public, private, and imported certificates for secure communications across the internet or within internal networks. The service supports both IPv4 and IPv6 on public endpoints and can secure singular domain names, multiple specific domain names, wildcard domains, or combinations of these. ACM automates certificate deployment with integrated AWS services, handles automatic renewals, and uses strong encryption and key management practices to protect private keys. Public certificates are trusted by browsers, operating systems, and mobile devices, while private certificates are used for internal applications and services.

Key Use Cases

- Securing websites and web applications: Deploy SSL/TLS certificates on Elastic Load Balancers, CloudFront distributions, and API Gateway to enable HTTPS communications
- Internal PKI for enterprise applications: Use private certificates signed by AWS Private CA for securing internal applications and services within corporate networks
- Container and Kubernetes workloads: Secure Amazon EKS pods and terminate TLS at Kubernetes Ingress or AWS load balancers using ACM certificates
- Nitro Enclaves security: Install certificates on EC2 instances with Nitro Enclaves for enhanced security applications requiring isolated certificate management

Features

- **Automatic Certificate Renewal**: ACM automatically renews certificates after 11 months of their 13-month validity period to ensure continuous security
- **Domain Validation**: Supports DNS, email, and HTTP validation methods to verify domain ownership before certificate issuance
- **Exportable Public Certificates**: Allows exporting public certificates for use on any workload including EC2 instances, containers, or on-premises hosts
- **AWS Service Integration**: Native integration with ELB, CloudFront, API Gateway, Elastic Beanstalk, App Runner, and other AWS services
- **Multiple Certificate Types**: Supports single domain, multi-domain, and wildcard certificates with various encryption algorithms (RSA 2048-bit, ECDSA 256/384-bit)
- **Private Certificate Authority Support**: Integration with AWS Private CA for issuing private certificates for internal applications

Value Proposition

AWS Certificate Manager eliminates the complexity and cost of purchasing, uploading, and renewing SSL/TLS certificates. Public certificates are provided at no additional cost for use with integrated AWS services, significantly reducing operational expenses compared to third-party certificate authorities:

- The service automates the entire certificate lifecycle including provisioning, deployment, and renewal, reducing manual errors and administrative overhead
- ACM's tight integration with AWS services ensures seamless certificate deployment and management
- The service provides enterprise-grade security with encrypted private key storage using AWS KMS, and supports both public and private PKI use cases
- Automatic renewal prevents service disruptions from expired certificates, while centralized management through the AWS console, CLI, and APIs simplifies operations across multiple certificates and domains

Service Integration

ACM integrates natively with multiple core AWS services for certificate deployment and management. Primary integrations include Elastic Load Balancer (ELB) for load balancer SSL termination, Amazon CloudFront for secure content delivery, and Amazon API Gateway for secure API endpoints:

- Additional integrations encompass AWS Elastic Beanstalk for application environments, AWS App Runner for containerized applications, Amazon EKS for Kubernetes workloads, and AWS Nitro Enclaves for enhanced security applications
- The service also works with Amazon Cognito for user pool configurations, AWS Amplify for custom domains, Amazon OpenSearch Service for Application Load Balancer provisioning, and AWS Network Firewall for TLS inspection
- AWS CloudFormation supports ACM certificates as template resources, enabling infrastructure-as-code deployments across all integrated services

Onboarding and Offboarding

Getting started with ACM requires signing into the AWS Management Console and navigating to the ACM console. Customers can request public certificates by specifying domain names and choosing validation methods (DNS, email, or HTTP):

- For DNS validation with Route 53, ACM can automatically create required CNAME records
- Domain ownership must be validated before certificate issuance
- Once validated, certificates can be deployed to integrated AWS services like load balancers or CloudFront distributions

- For exportable certificates, customers can retrieve the certificate and private key for use on any workload
- Private certificates require setting up AWS Private CA first
- The service provides AWS CLI and API access for programmatic management, and supports Infrastructure-as-Code through CloudFormation templates

Getting Started Guide: https://docs.aws.amazon.com/acm/latest/userguide/gs.html

Data Removal

Certificate removal from ACM involves deleting certificates through the console, CLI, or API. When certificates are deleted, they are immediately removed from ACM and can no longer be used with AWS services. For certificates deployed to integrated services, customers must first remove them from those services before deletion. Private keys stored in ACM are securely deleted using AWS KMS key management. Imported certificates can be deleted without affecting the original certificate authority. Certificate transparency logs may retain public certificate information as they are append-only systems managed by third parties.

Backup/Restore and Disaster Recovery

ACM certificates are regional resources that must be requested or imported separately in each region. The service provides built-in resilience through multiple Availability Zones within each region. For disaster recovery, customers should provision certificates in multiple regions and implement cross-region failover strategies. ACM does not provide traditional backup functionality as certificates can be re-issued or imported as needed. Private CA certificates can be backed up through AWS Private CA's disaster recovery capabilities including cross-region subordinate CA deployment strategies.

Backup Capabilities

ACM does not offer traditional backup capabilities as certificates are managed resources that can be re-provisioned. The service maintains certificate metadata and can re-issue certificates upon request:

- For exportable certificates, customers can export and securely store certificates externally
- ACM does not integrate with AWS Backup service as certificates are not considered data requiring backup
- Certificate renewal maintains continuity without requiring backup restoration processes

Disaster Recovery

ACM supports disaster recovery through regional redundancy and certificate re-provisioning capabilities. Certificates are regional resources requiring separate provisioning in each region for multi-region DR strategies:

- The service does not integrate with AWS Elastic Disaster Recovery as certificates can be quickly re-issued or imported during recovery scenarios
- For comprehensive DR, customers should provision certificates in multiple regions and implement automated failover mechanisms using services like Route 53 health checks

Recovery Objectives

ACM supports rapid certificate provisioning to meet Recovery Time Objectives (RTO), with public certificates issued within seconds after domain validation. Recovery Point Objectives (RPO) are minimal as certificate metadata is maintained by AWS. For disaster recovery scenarios, certificates can be quickly re-issued or imported in alternate regions. Automated renewal ensures certificates remain valid during extended outages. Integration with AWS services enables rapid restoration of secure communications during recovery operations.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/acm/latest/userguide/acm-limits.html

FAQ: https://aws.amazon.com/certificate-manager/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/acm/latest/userguide/acm-overview.html

User Guide: https://docs.aws.amazon.com/acm/latest/userguide/gs.html

API Reference: https://docs.aws.amazon.com/acm/latest/APIReference/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/certificate-manager/pricing/

Cost Optimization

ACM provides significant cost optimization through free public certificates for integrated AWS services, eliminating annual certificate authority fees that can cost hundreds of

dollars per certificate. Exportable public certificates cost only $15 per FQDN and $149 per wildcard, significantly below market rates:

- The service includes 10,000 free certificate export API calls monthly
- Automatic renewal eliminates manual certificate management costs and prevents expensive service outages from expired certificates
- Wildcard certificates can secure multiple subdomains under a single certificate, reducing the total number of certificates needed
- Import capabilities allow using existing certificates without additional ACM charges
- The service's integration with AWS services eliminates third-party SSL proxy costs and reduces infrastructure complexity

Savings Considerations

Primary cost savings come from eliminating traditional certificate authority fees, which typically range from $50-500+ annually per certificate. Organizations can save thousands of dollars annually by replacing third-party certificates with free ACM public certificates:

- Automatic renewal prevents costly service disruptions and eliminates manual certificate management labor costs
- Wildcard certificates provide economies of scale for organizations with many subdomains
- The service reduces operational costs through centralized certificate management and automated deployment
- Integration with AWS services eliminates need for dedicated certificate management infrastructure
- For exportable certificates, ACM's pricing is substantially lower than traditional enterprise certificate authorities while providing the same security and trust levels
- Organizations moving from on-premises certificate management can realize significant cost reductions in both licensing and operational expenses

Service Name

AWS Clean Rooms

Service Overview

AWS Clean Rooms is a secure collaboration workspace that enables organizations to analyze and collaborate on their collective datasets with partners without revealing underlying data. The service allows users to create clean rooms in minutes, invite partners, link data resources, and configure privacy-enhancing controls. Organizations can run SQL queries, PySpark jobs, and machine learning models on combined datasets while maintaining data privacy through built-in access controls, analysis rules, cryptographic computing, audit logs, and differential privacy. Data remains in its original location without requiring movement to another cloud provider.

Key Use Cases

- Advertising and Marketing: Improve advertiser reach and publisher monetization through secure collaboration on campaign data and audience insights
- Machine Learning and Analytics: Train custom ML models and create lookalike segments using collective datasets for fraud detection, campaign optimization, and medical research
- Entity Resolution: Match and link customer records from various sources using identity resolution partners like LiveRamp and Neustar to generate common IDs
- Cross-industry Data Collaboration: Enable secure data sharing for research and development, investment decisions, and multi-party analytics across healthcare, finance, and other regulated industries

Features

- **Multi-party Collaboration**: Create collaborations with multiple organizations and configure member roles for secure data sharing
- **SQL and PySpark Analytics**: Run SQL queries and PySpark jobs on collective datasets with Spark SQL support and configurable compute sizes
- **Privacy-Enhancing Controls**: Implement fine-grained analysis rules, differential privacy, and cryptographic computing to protect sensitive data
- **Cross-Cloud Support**: Collaborate with datasets from multiple clouds including AWS, Snowflake, and Amazon Athena
- **Machine Learning Capabilities**: AWS Clean Rooms ML for custom modeling, lookalike modeling, and synthetic dataset generation
- **Entity Resolution**: AWS Entity Resolution integration for matching customer records across different data sources

- **Cross-Region Collaboration**: Support for collaboration with data sources stored in different AWS and Snowflake regions
- **Analysis Templates**: Pre-approved SQL and Python queries that can be reused across collaborations

Value Proposition

AWS Clean Rooms provides unique value through its comprehensive privacy-preserving data collaboration platform that eliminates the need to move or share raw data. Organizations benefit from IDC MarketScape recognition as a leader in worldwide data clean room technology, extensive privacy controls including differential privacy and cryptographic computing, seamless integration with existing AWS services and third-party platforms like Snowflake, and the ability to collaborate with thousands of companies already using AWS. The service offers flexibility through multiple analytics engines (SQL, PySpark), supports both custom and lookalike ML modeling, and provides cost-effective collaboration where compute costs can be shared between partners.

Service Integration

AWS Clean Rooms integrates deeply with core AWS services to provide comprehensive data collaboration capabilities. Key integrations include Amazon S3 for data storage, AWS Glue Data Catalog for data preparation and cataloging, Amazon Athena for serverless querying, AWS CloudFormation for infrastructure as code deployment, AWS CloudTrail for audit logging, AWS Entity Resolution for identity matching, AWS Secrets Manager for secure credential storage, Amazon EventBridge for event-driven workflows, and AWS KMS for encryption key management. The service also works with Amazon SageMaker for ML model deployment, Amazon QuickSight for visualization, and AWS Backup for data protection, creating a complete data collaboration ecosystem within the AWS cloud.

Onboarding and Offboarding

Customers get started with AWS Clean Rooms by first signing up for an AWS account and setting up necessary service roles for AWS Clean Rooms and AWS Clean Rooms ML using the AWS Management Console, CLI, or CloudFormation. The onboarding process involves creating a collaboration or joining an existing one through membership creation, configuring data sources by linking AWS Glue tables, Amazon S3 data, or Snowflake sources, setting up analysis rules and privacy controls, and optionally creating analysis templates for pre-approved queries. The service provides guided workflows through AWS Clean Rooms Solutions to simplify data preparation and collaboration setup, with comprehensive documentation and expert support available.

Getting Started Guide: https://docs.aws.amazon.com/clean-rooms/latest/userguide/setting-up.html

Data Removal

During offboarding, collaboration creators and members can delete their collaborations, memberships, and associated resources through the AWS Clean Rooms console or API. Data removal involves deleting configured tables, analysis templates, and any query results stored in Amazon S3. Organizations retain control over their original data sources, which remain in their own AWS accounts and are never moved to AWS Clean Rooms infrastructure, ensuring complete data sovereignty during the offboarding process.

Backup/Restore and Disaster Recovery

AWS Clean Rooms leverages underlying AWS service backup capabilities since data remains in original locations like Amazon S3 and AWS Glue. Organizations can use AWS Backup for centralized backup management, Amazon S3 versioning and replication for data protection, and cross-region data sources for geographic redundancy. The service metadata and configurations can be recreated using AWS CloudFormation templates for disaster recovery scenarios.

Backup Capabilities

AWS Clean Rooms does not provide native backup capabilities as it operates on data stored in external sources. Backup responsibilities lie with the underlying data storage services like Amazon S3, which offers versioning, replication, and integration with AWS Backup. Organizations should implement backup strategies for their source data systems and use AWS CloudFormation templates to backup collaboration configurations and analysis templates.

Disaster Recovery

AWS Clean Rooms supports disaster recovery through its distributed architecture across multiple AWS regions and integration with resilient AWS services. The service supports cross-region collaboration, allowing data sources in different regions to participate in the same collaboration. Recovery strategies can leverage AWS CloudFormation for infrastructure recreation, Amazon S3 cross-region replication for data availability, and the service's stateless query execution model for rapid recovery.

Recovery Objectives

AWS Clean Rooms helps customers meet RPO and RTO objectives through its integration with highly available AWS services and cross-region support. The service's architecture allows for quick collaboration recreation using CloudFormation templates, while underlying data services like Amazon S3 provide sub-second RPO through cross-region replication. RTO objectives can be met through automated failover to alternate regions and the service's ability to quickly recreate collaboration environments without data movement requirements.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/clean-rooms/latest/userguide/quotas.html

FAQ: https://aws.amazon.com/clean-rooms/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/clean-rooms/latest/userguide/what-is.html

User Guide: https://docs.aws.amazon.com/clean-rooms/latest/userguide/

API Reference: https://docs.aws.amazon.com/clean-rooms/latest/apireference/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/clean-rooms/pricing/

Cost Optimization

AWS Clean Rooms offers unique cost optimization through its pay-per-query model based on Clean Rooms Processing Units (CRPUs), allowing organizations to pay only for actual compute usage. Cost optimization features include configurable compute sizes for SQL queries, ability to cache tables for complex queries to reduce costs, customizable Spark properties for performance tuning, and shared cost model where collaboration creators can designate which member pays for compute costs. The service also supports advanced configurations to optimize SQL performance and reduce processing time, directly impacting costs.

Savings Considerations

Main cost savings come from eliminating data movement and duplication costs since data remains in original locations, shared compute cost model where partners can split collaboration expenses, efficient query execution through optimized Spark SQL engine and configurable compute resources, and elimination of traditional ETL processes for data collaboration. Organizations save on infrastructure costs by leveraging existing AWS data storage and avoiding dedicated data clean room infrastructure. The pay-per-use model ensures costs scale with actual usage rather than requiring upfront capacity planning or reserved resources.

Service Name

AWS CloudFormation

Service Overview

AWS CloudFormation is a service that helps model and set up AWS resources using infrastructure as code. It allows users to create templates in JSON or YAML format that describe all desired AWS resources and their properties. CloudFormation provisions and configures these resources automatically, eliminating the need for manual creation and configuration. The service manages resources as collections called stacks, which can be created, updated, and deleted as single units. CloudFormation handles dependencies between resources and provides change sets for previewing modifications before deployment, with automatic rollback capabilities if errors occur during provisioning.

Key Use Cases

- Infrastructure as Code: Define and version control AWS infrastructure using JSON or YAML templates for consistent, repeatable deployments across environments
- Multi-region deployment: Replicate application infrastructure across multiple AWS regions using the same templates for disaster recovery and global availability
- Environment management: Create and manage development, staging, and production environments with consistent configurations while enabling easy cleanup through stack deletion
- DevOps automation: Integrate with CI/CD pipelines to automate infrastructure provisioning and updates as part of application deployment workflows
- Compliance and governance: Implement standardized infrastructure configurations with built-in drift detection and policy enforcement across multiple accounts using StackSets

Features

- **Template-based Infrastructure**: Define infrastructure using JSON or YAML templates with support for parameters, mappings, conditions, and outputs
- **Stack Management**: Deploy and manage resources as cohesive units called stacks with full lifecycle management capabilities
- **Change Sets**: Preview proposed changes before deployment to understand the impact of template modifications
- **Automatic Rollback**: Automatically rollback to previous working state if stack creation or updates fail
- **Drift Detection**: Detect when resources have been modified outside of CloudFormation and remediate configuration drift

- **StackSets**: Deploy stacks across multiple AWS accounts and regions with centralized management and governance
- **Custom Resources**: Extend CloudFormation capabilities with custom provisioning logic for third-party or custom AWS resources
- **Import Existing Resources**: Bring existing AWS resources under CloudFormation management without recreating them

Value Proposition

CloudFormation provides native AWS integration with comprehensive support for all AWS services and deep understanding of resource dependencies and relationships. It offers no additional cost for the service itself, with customers paying only for underlying AWS resources:

- The service includes advanced features like pre-deployment validation, operation tracking, and simplified troubleshooting capabilities
- CloudFormation integrates seamlessly with AWS developer tools, supports Infrastructure Composer for visual template design, and provides extensive sample templates
- It offers enterprise-grade capabilities including StackSets for multi-account deployments, drift detection for compliance, and automatic rollback for reliability, making it the authoritative choice for AWS infrastructure automation

Service Integration

CloudFormation integrates natively with virtually all AWS services including compute services like EC2, Lambda, and ECS; storage services like S3, EBS, and EFS; database services like RDS, DynamoDB, and ElastiCache; networking services like VPC, Route53, and Load Balancers; and security services like IAM, KMS, and Secrets Manager. It works closely with AWS developer tools including CodePipeline, CodeBuild, and CodeDeploy for CI/CD integration:

- CloudFormation supports AWS Config for compliance monitoring, CloudTrail for audit logging, and SNS for event notifications
- The service extends through AWS CDK and SAM for higher-level abstractions, and integrates with AWS Backup for stack-level backup and recovery capabilities

Onboarding and Offboarding

Customers can start with CloudFormation by creating an AWS account and accessing the service through the AWS Management Console, CLI, or SDKs. The Getting Started guide provides step-by-step instructions for creating the first stack using sample templates:

- AWS offers guided learning resources including workshops and sample templates organized by service and use case

- Users can begin with simple templates and progressively adopt advanced features like parameters, conditions, and nested stacks
- For visual learners, AWS Infrastructure Composer provides a drag-and-drop interface for template creation
- The service includes extensive documentation, API references, and best practice guides to accelerate adoption and ensure successful implementation

Getting Started Guide: https://docs.aws.amazon.com/AWSCloudFormation/latest/UserGuide/GettingStarted.html

Data Removal

CloudFormation provides straightforward data removal through stack deletion, which automatically removes all resources created by the stack in the correct order while respecting dependencies. Resources with DeletionPolicy set to Retain or Snapshot are preserved during stack deletion. For complete offboarding, customers can delete all CloudFormation stacks, remove saved templates from S3 buckets, and delete any retained resources manually. Stack termination protection prevents accidental deletion of critical stacks, and customers can export stack outputs before deletion for migration purposes.

Backup/Restore and Disaster Recovery

CloudFormation provides backup capabilities through integration with AWS Backup, which can create composite recovery points for entire CloudFormation stacks including templates and associated resources. Templates themselves serve as infrastructure blueprints that can be stored in version control systems and S3 for disaster recovery. The service supports cross-region template replication for disaster recovery scenarios and enables quick infrastructure recreation in alternate regions using the same templates, providing infrastructure-level disaster recovery capabilities.

Backup Capabilities

CloudFormation integrates with AWS Backup to create composite backups of entire stacks, including both the CloudFormation template and recovery points for supported resources within the stack. These composite recovery points group related backup data together, enabling restoration of complete application environments. Templates can be versioned and stored in S3 with cross-region replication, and the service maintains stack event history for operational insights and troubleshooting purposes.

Disaster Recovery

CloudFormation supports disaster recovery through template-based infrastructure recreation in alternate regions, enabling rapid recovery of complete environments. The service works with AWS Backup for stack-level recovery and supports cross-region

StackSets for deploying identical infrastructure across multiple regions. While CloudFormation doesn't directly integrate with AWS Elastic Disaster Recovery, it provides the foundation for disaster recovery strategies by enabling consistent infrastructure deployment across regions and availability zones.

Recovery Objectives

CloudFormation helps meet RPO objectives by enabling point-in-time infrastructure snapshots through templates and integration with AWS Backup for application data. For RTO objectives, the service enables rapid infrastructure recreation through automated template deployment, with stack creation times typically measured in minutes depending on resource complexity. Pre-validated templates and change sets reduce deployment failures, while cross-region template replication enables quick disaster recovery site activation, supporting both conservative and aggressive RTO/RPO targets based on architecture design.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/AWSCloudFormation/latest/UserGuide/cloudformation-limits.html

FAQ: https://aws.amazon.com/cloudformation/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/AWSCloudFormation/latest/UserGuide/default.html

User Guide: https://docs.aws.amazon.com/AWSCloudFormation/latest/UserGuide/Welcome.html

API Reference: https://docs.aws.amazon.com/AWSCloudFormation/latest/APIReference/

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/cloudformation/pricing/

Cost Optimization

CloudFormation itself is free, with customers paying only for underlying AWS resources, making it inherently cost-effective. The service enables cost optimization through infrastructure automation, eliminating manual provisioning errors that can lead to resource waste:

- Templates can include cost-optimized resource configurations like right-sized instances, appropriate storage classes, and scheduled resource shutdown
- Stack-level resource management enables easy cleanup of entire environments when no longer needed
- Integration with AWS Cost Explorer and Budgets allows for cost monitoring at the stack level, while parameterized templates enable cost-conscious resource selection based on environment requirements

Savings Considerations

Primary cost savings come from infrastructure standardization and automation, reducing operational overhead and preventing resource sprawl. CloudFormation enables environment-specific resource sizing through parameters, allowing development environments to use smaller, less expensive resources:

- Automated stack deletion prevents forgotten resources from incurring ongoing costs, while drift detection helps identify unauthorized resource modifications that may impact costs
- Template reuse across environments reduces development time and ensures consistent, cost-optimized configurations
- The service supports implementation of cost governance policies through standardized templates and enables integration with AWS cost management tools for ongoing optimization and monitoring

Service Name

AWS CloudHSM

Service Overview

AWS CloudHSM is a cloud-based hardware security module (HSM) service that enables customers to generate and manage their own encryption keys using dedicated, FIPS 140-2 Level 3 validated HSMs. The service provides complete control over high-availability HSMs in the AWS Cloud with low-latency access and secure cryptographic operations. CloudHSM combines the benefits of AWS cloud scalability with the security of hardware security modules, offering single-tenant access to computing devices that process cryptographic operations and provide secure storage for cryptographic keys. The service automates HSM management including backups, provisioning, configuration, and maintenance while ensuring customers retain full control over their encryption keys and cryptographic operations.

Key Use Cases

- Regulatory compliance: Meet corporate, contractual, and regulatory compliance requirements for data security using dedicated HSMs that comply with PCI DSS, PCI PIN, and PCI-3DS standards
- Database encryption: Enable Transparent Data Encryption (TDE) for databases by storing TDE master encryption keys in HSMs to encrypt database files with enhanced security
- SSL/TLS processing offload: Offload SSL/TLS processing for web servers to reduce computational burden and provide additional security for high-traffic applications
- Certificate Authority (CA) operations: Store and manage private keys for certificate authorities to maintain trust and perform cryptographic signing operations for PKI infrastructure
- Document signing and verification: Create asymmetric public and private key pairs for digitally signing and verifying documents to ensure authenticity and integrity
- Message authentication: Securely create and manage symmetric keys for CMACs and HMACs to authenticate messages and ensure message integrity
- Code and firmware signing: Sign applications and firmware using private keys stored in HSMs to establish trust and prevent unauthorized code execution
- Key management integration: Combine with AWS KMS to store key material in a single-tenant environment while leveraging AWS KMS benefits for integrated AWS services
- Random number generation: Securely generate cryptographically strong random numbers in HSMs for creating encryption keys and other security tokens

Features

- **FIPS Validated HSMs**: FIPS 140-2 Level 3 and FIPS 140-3 Level 3 validated hardware security modules for clusters in FIPS mode, with non-FIPS mode available for broader algorithm support
- **Industry-Standard APIs**: Support for PKCS#11, Java Cryptography Extensions (JCE), Microsoft CryptoNG (CNG), and Key Storage Provider (KSP) APIs for seamless application integration
- **High Availability Clustering**: Multi-AZ cluster deployment with automatic synchronization and load balancing across multiple HSMs for 99.95% uptime SLA
- **Automatic Backup Management**: Automated daily backups with configurable retention policies, cross-region backup copying, and cluster restore capabilities
- **Single-Tenant Architecture**: Dedicated HSM instances with complete isolation from other customers and full customer control over keys and cryptographic operations
- **Scalable Performance**: Up to 10,000 cryptographic operations per second with support for up to 28 HSMs per cluster and storage for up to 16,666 keys
- **Client SDK Integration**: Comprehensive client SDKs for multiple platforms including Linux and Windows with CloudHSM CLI for cluster management
- **Third-Party Integrations**: Native support for Oracle TDE, SSL offload, Windows CA, and integration with third-party vendors like HashiCorp Vault and PrimeKey EJBCA

Value Proposition

AWS CloudHSM provides customers with dedicated, single-tenant hardware security modules that offer complete control over cryptographic operations while leveraging AWS cloud benefits. Unlike managed services, CloudHSM gives customers full ownership of encryption keys with end-to-end encryption invisible to AWS:

- The service offers superior security through FIPS 140-2/140-3 Level 3 validation, ensuring compliance with stringent regulatory requirements
- Key advantages include seamless migration support for existing PKCS#11, JCE, and CNG applications, automatic high availability across multiple AZs, and integration with industry-standard cryptographic APIs
- The fully managed infrastructure eliminates hardware procurement, maintenance, and patching overhead while providing elastic scalability and global availability across multiple AWS regions

Service Integration

AWS CloudHSM integrates with Amazon VPC for network isolation and secure connectivity to EC2 instances running client applications. Integration with AWS KMS enables custom key stores, allowing customers to combine CloudHSM's single-tenant security with KMS's seamless AWS service integration for services like S3, EBS, and RDS:

- CloudWatch provides monitoring and alerting for HSM health and performance metrics
- CloudTrail logs all AWS CloudHSM API calls for audit and compliance purposes
- The service works with AWS Direct Connect and VPN for hybrid connectivity, enabling secure communication between on-premises applications and cloud-hosted HSMs
- Integration with IAM provides access control for CloudHSM management operations, while AWS Backup can be used for additional backup orchestration across AWS services

Onboarding and Offboarding

Getting started with AWS CloudHSM involves creating a VPC and subnets across multiple availability zones, then creating a CloudHSM cluster through the console, CLI, or API. Customers must initialize the cluster with a self-signed certificate, add HSMs to achieve desired availability and performance, and install the CloudHSM client software on EC2 instances:

- The process includes configuring security groups for HSM communication, activating the cluster using CloudHSM CLI, creating HSM users (crypto officers and crypto users), and integrating applications using supported APIs like PKCS#11 or JCE
- AWS provides comprehensive documentation, sample code, and getting started guides
- Setup typically requires network configuration knowledge and understanding of cryptographic concepts, but AWS offers professional services support for complex implementations

Getting Started Guide: https://docs.aws.amazon.com/cloudhsm/latest/userguide/getting-started.html

Data Removal

To offboard from AWS CloudHSM, customers must export any required keys using supported APIs before deleting HSMs and clusters. The process involves stopping client applications, removing HSMs from clusters using the console or CLI, and deleting the cluster itself. All HSMs must be removed before cluster deletion. Customer data in HSMs is automatically destroyed when HSMs are deleted due to tamper-resistant hardware design. Backup data can be explicitly deleted through on-demand backup deletion features, though backups are retained for 7 days after deletion requests for recovery purposes.

Backup/Restore and Disaster Recovery

AWS CloudHSM automatically creates encrypted backups of clusters at least once every 24 hours, including all user data, key material, certificates, and HSM configuration. Backups can be copied across AWS regions for disaster recovery and used to create new clusters in different regions. The service supports configurable backup retention policies with automatic purging of expired backups. Customers can restore entire clusters from backups and force backup creation by adding/removing HSMs. Multi-AZ cluster deployment provides high availability with automatic failover between HSMs in different availability zones.

Backup Capabilities

AWS CloudHSM includes native automated backup capabilities with daily encrypted backups stored securely by AWS. The service provides managed backup retention policies, on-demand backup deletion through AWS SDK and CLI, and cross-region backup copying for geographic distribution:

- While AWS CloudHSM does not directly integrate with AWS Backup service, its comprehensive built-in backup features provide robust data protection
- Backups include complete cluster state including keys, users, and configurations, enabling full cluster restoration

Disaster Recovery

AWS CloudHSM supports disaster recovery through multi-AZ cluster deployment, cross-region backup copying, and the ability to create clusters from backups in different regions. The service does not directly integrate with AWS Elastic Disaster Recovery, but provides comprehensive DR capabilities through its native backup and clustering features. Customers can implement cross-region disaster recovery by copying backups to secondary regions and creating standby clusters when needed.

Recovery Objectives

AWS CloudHSM supports low RTO objectives through multi-AZ high availability clustering with automatic failover between HSMs in different availability zones. RPO objectives are met through automated daily backups with the ability to force more frequent backups as needed. Cross-region backup copying enables geographic disaster recovery. The 99.95% uptime SLA for multi-AZ clusters helps meet stringent availability requirements. Customers can achieve near-zero RTO within a region through cluster redundancy and low RPO through frequent backup operations.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/cloudhsm/latest/userguide/limits.html

FAQ: https://aws.amazon.com/cloudhsm/faqs/

Technical Requirements

Service Documentation:
https://docs.aws.amazon.com/cloudhsm/latest/userguide/introduction.html

User Guide: https://docs.aws.amazon.com/cloudhsm/latest/userguide/introduction.html

API Reference:
https://docs.aws.amazon.com/cloudhsm/latest/APIReference/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/cloudhsm/pricing/

Cost Optimization

AWS CloudHSM offers unique cost optimization through pay-as-you-go hourly pricing with no upfront costs or long-term commitments. Customers can dynamically scale HSM capacity by adding or removing HSMs based on workload demands:

- The service eliminates capital expenses for HSM hardware, maintenance, and facilities
- Costs are optimized through managed backup retention policies that automatically purge expired backups, and on-demand backup deletion capabilities
- Regional deployment flexibility allows customers to choose cost-effective regions while maintaining compliance requirements
- The fully managed service reduces operational overhead and staffing costs compared to on-premises HSM management

Savings Considerations

Primary cost savings include elimination of hardware procurement, maintenance, and replacement costs for physical HSMs. Customers avoid facility requirements including secure data centers, power, cooling, and physical security infrastructure:

- The managed service reduces operational staffing needs for HSM maintenance, firmware updates, and hardware monitoring
- Elastic scaling prevents over-provisioning by allowing HSM capacity adjustment based on actual demand
- Automated backup management reduces storage costs through intelligent retention policies
- No reserved instance pricing is offered, but the pay-per-hour model provides cost predictability

- Cross-region backup copying eliminates need for separate disaster recovery infrastructure investments

Service Name

AWS CloudTrail

Service Overview

AWS CloudTrail is a service that enables operational and risk auditing, governance, and compliance of AWS accounts. It records actions taken by users, roles, or AWS services as events, providing a comprehensive audit trail of account activity. CloudTrail captures API calls made through the AWS Management Console, AWS CLI, and AWS SDKs, delivering these events to Amazon S3 buckets, CloudWatch Logs, or EventBridge. The service offers three ways to record events: Event History (90 days of management events at no cost), CloudTrail Lake (managed data lake for long-term storage and analysis), and Trails (configurable event delivery to S3).

Key Use Cases

- Security analysis and incident response: Monitor and investigate suspicious activities, unauthorized access, and security breaches across AWS resources
- Compliance auditing: Meet regulatory requirements by maintaining detailed logs of all AWS account activities for audit trails
- Operational troubleshooting: Debug issues by analyzing API call patterns, error rates, and resource interactions to identify root causes

Features

- **Event History**: Always-on recording of the past 90 days of management events with search, view, and download capabilities at no additional cost
- **CloudTrail Lake**: Managed data lake for capturing, storing, and analyzing events with retention up to 10 years, SQL-based querying, and support for multiple data sources
- **Trails**: Configurable event delivery to S3 buckets with support for management events, data events, network activity events, and multi-region logging
- **CloudTrail Insights**: Machine learning-powered anomaly detection to identify unusual API call patterns and error rates
- **Multi-region and Multi-account Support**: Centralized logging across all AWS regions and organization accounts with consolidated event delivery

Value Proposition

CloudTrail provides comprehensive, native AWS audit logging without requiring additional infrastructure setup or management. It offers automatic encryption, immutable event records, and seamless integration with other AWS security services like GuardDuty and Security Hub:

- The service scales automatically with AWS usage, provides real-time monitoring capabilities, and offers flexible retention options from 90 days to 10 years
- Unlike third-party logging solutions, CloudTrail captures all AWS service interactions natively, requires no agent installation, and provides cost-effective storage with multiple delivery options including S3, CloudWatch Logs, and EventBridge

Service Integration

CloudTrail integrates deeply with Amazon S3 for log file storage, Amazon CloudWatch for real-time monitoring and alerting, and Amazon EventBridge for event-driven architectures. It works with AWS KMS for encryption of log files and event data stores, AWS Organizations for centralized multi-account logging, and AWS IAM for access control:

- Additional integrations include Amazon SNS for delivery notifications, Amazon Athena for log analysis, AWS Config for configuration tracking, Amazon GuardDuty for threat detection, and AWS Security Hub for centralized security findings
- CloudTrail also supports integration with third-party SIEM solutions and custom applications through its APIs

Onboarding and Offboarding

CloudTrail is automatically enabled for all AWS accounts with Event History providing 90 days of management event logs at no cost. To get started with advanced features, customers can create trails through the AWS Management Console, AWS CLI, or APIs to deliver events to S3 buckets:

- CloudTrail Lake setup involves creating event data stores with chosen retention periods and pricing options
- The service provides managed policies (AWSCloudTrail_FullAccess and AWSCloudTrail_ReadOnlyAccess) for permissions management
- Getting started requires minimal configuration - simply specify an S3 bucket for trail delivery or create an event data store for Lake functionality

Getting Started Guide:
https://docs.aws.amazon.com/awscloudtrail/latest/userguide/cloudtrail-user-guide.html

Data Removal

Offboarding involves deleting trails, event data stores, and channels through the console or APIs. Deleting trails stops logging but doesn't remove existing log files in S3 buckets. Event data stores enter a PENDING_DELETION state for 7 days before permanent deletion, during which no queries or operations are possible. Customers must manually delete log files from S3 buckets and CloudWatch Logs if complete data removal is required.

Backup/Restore and Disaster Recovery

CloudTrail provides immutable audit logs stored in S3 with optional cross-region replication for disaster recovery. Event data stores offer durable storage with automatic backup across multiple Availability Zones. The service supports log file integrity validation using digest files to ensure data hasn't been tampered with. CloudTrail integrates with AWS Backup for centralized backup management of related resources.

Backup Capabilities

CloudTrail logs are automatically stored with high durability in S3 and replicated across multiple Availability Zones. Event data stores provide immutable storage with configurable retention periods:

- The service doesn't integrate directly with AWS Backup but supports S3 cross-region replication for additional protection
- Log file integrity validation ensures audit trail authenticity

Disaster Recovery

CloudTrail supports disaster recovery through multi-region trail configuration and S3 cross-region replication. Event data stores can be configured for multi-region logging with automatic failover capabilities. The service doesn't directly integrate with AWS Elastic Disaster Recovery but provides the audit trail necessary for compliance during disaster recovery scenarios.

Recovery Objectives

CloudTrail helps meet RPO objectives through real-time event logging with minimal delay (typically within 15 minutes). For RTO objectives, multi-region trails and event data stores provide rapid access to audit logs even during regional outages. Cross-region replication and immutable storage ensure audit continuity during disaster recovery scenarios.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/awscloudtrail/latest/userguide/WhatIsCloudTrail-Limits.html

FAQ: https://aws.amazon.com/cloudtrail/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/awscloudtrail/latest/userguide/cloudtrail-user-guide.html

User Guide: https://docs.aws.amazon.com/awscloudtrail/latest/userguide/cloudtrail-user-guide.html

API Reference: https://docs.aws.amazon.com/awscloudtrail/latest/APIReference/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/cloudtrail/pricing/

Cost Optimization

CloudTrail offers several cost optimization options including advanced event selectors to filter unnecessary events, choosing appropriate retention periods (1-year extendable vs 7-year for CloudTrail Lake), and using the free 90-day Event History for basic needs. Customers can optimize costs by creating single multi-region trails instead of multiple regional trails, using data event aggregation to reduce volume, and selecting the most cost-effective pricing tier based on ingestion patterns. S3 lifecycle policies can reduce storage costs for long-term log retention, and careful selection of event types prevents unnecessary logging charges.

Savings Considerations

Major cost savings come from avoiding duplicate trail creation (first copy of management events per region is free), using advanced event selectors to log only necessary resources, and choosing CloudTrail Lake one-year extendable retention for variable workloads. Organizations can save by using organization trails instead of individual account trails:

- For high-volume environments, aggregating data events and using appropriate retention periods based on compliance requirements provides significant savings
- Leveraging the free Event History tier for basic monitoring needs and upgrading to paid features only when advanced capabilities are required optimizes costs effectively

Service Name

AWS Cloud Map

Service Overview

AWS Cloud Map is a fully managed cloud resource discovery service that enables applications to automatically discover and connect to backend services and resources using logical names. It serves as a service registry that maps meaningful names to dynamic cloud resources like Amazon EC2 instances, Amazon ECS tasks, Amazon DynamoDB tables, Amazon SQS queues, and other AWS services. Applications can locate these resources through DNS queries or API calls, with Cloud Map only serving healthy resources. The service simplifies microservices architecture by providing a centralized way to track and connect to distributed application components as they dynamically change locations and endpoints.

Key Use Cases

- Service discovery for microservices: Enable microservices to automatically discover and connect to each other in dynamic containerized environments using Amazon ECS and Amazon EKS
- Dynamic application resource mapping: Map logical names to frequently changing cloud resources like load balancers, databases, and compute instances that scale up and down
- Health-aware service routing: Automatically route traffic only to healthy service instances by integrating with Amazon Route 53 health checks and custom health checkers
- Cross-account service discovery: Share service registries across multiple AWS accounts using AWS Resource Access Manager for centralized service management in multi-account architectures

Features

- **Namespace Management**: Create HTTP, private DNS, and public DNS namespaces to group and organize application services with different discovery methods
- **Service Registration**: Register and manage service instances with custom attributes for filtering and discovery using RegisterInstance API
- **Health Checking**: Automatic health monitoring using Amazon Route 53 health checks or custom health checkers to serve only healthy resources
- **Discovery Methods**: Support for both DNS-based discovery and API-based discovery using DiscoverInstances for flexible service location

- **Custom Attributes**: Assign custom metadata attributes to services and instances for advanced filtering and routing decisions
- **Service-Level Attributes**: Store and access service-level information like traffic weights and configuration directly at the service level
- **Cross-Account Sharing**: Share Cloud Map namespaces across AWS accounts using AWS Resource Access Manager for centralized service registries
- **Container Integration**: Deep integration with Amazon ECS and Amazon EKS for automatic service registration and deregistration

Value Proposition

AWS Cloud Map provides significant advantages over traditional service discovery solutions by offering a fully managed, highly available service that eliminates the operational overhead of maintaining service registries. Unlike static configuration files or custom-built discovery solutions, Cloud Map automatically handles service registration, health checking, and provides multiple discovery mechanisms (DNS and API) in a single service:

- The deep integration with AWS container services like ECS and EKS enables automatic service lifecycle management, reducing manual effort and potential errors
- Its support for cross-account sharing through AWS RAM makes it ideal for enterprise multi-account architectures, while the ability to use both DNS queries and API calls provides flexibility for different application patterns and requirements

Service Integration

AWS Cloud Map integrates seamlessly with core AWS services to provide comprehensive service discovery capabilities. It works natively with Amazon ECS and Amazon EKS for automatic container service registration and deregistration, and with AWS Fargate for serverless container deployments:

- The service leverages Amazon Route 53 for DNS-based discovery and health checking, automatically creating hosted zones for DNS namespaces
- Integration with AWS App Mesh enables advanced service mesh capabilities for microservices communication and traffic management
- Cloud Map also works with AWS Identity and Access Management (IAM) for fine-grained access control, AWS Resource Access Manager (RAM) for cross-account namespace sharing, and supports various AWS resources including Amazon EC2 instances, Amazon DynamoDB tables, Amazon S3 buckets, Amazon SQS queues, and Amazon API Gateway APIs as discoverable services

Onboarding and Offboarding

Getting started with AWS Cloud Map involves creating a namespace to group services, then creating services within that namespace, and finally registering service instances. Users can access Cloud Map through the AWS Management Console, AWS CLI, SDKs, or REST API:

- The process begins by choosing the appropriate namespace type (HTTP for API-only discovery, private DNS for VPC-based DNS queries, or public DNS for internet-accessible services), followed by creating service templates that define health check configurations and DNS settings
- Applications then register instances using the RegisterInstance API and discover services using either DNS queries or the DiscoverInstances API
- AWS provides comprehensive tutorials, code examples, and CLI-based getting started guides to help users implement service discovery patterns quickly and effectively

Getting Started Guide: https://docs.aws.amazon.com/cloud-map/latest/dg/tutorials.html

Data Removal

To offboard from AWS Cloud Map, customers must deregister all service instances using the DeregisterInstance API, delete all services using DeleteService, and finally delete namespaces using DeleteNamespace. When deleting DNS-based namespaces, the corresponding Amazon Route 53 hosted zones are automatically removed. All registered service data, custom attributes, and health check configurations are permanently deleted. The deregistration process should be performed in reverse order of creation to avoid dependencies. AWS provides CLI commands and API operations to automate the cleanup process for bulk operations.

Backup/Restore and Disaster Recovery

AWS Cloud Map does not provide traditional backup and restore capabilities as it serves as a dynamic service registry rather than a data store. The service configuration and registered instances are designed to be ephemeral and recreated automatically by applications. However, Cloud Map itself is a highly available, fully managed service that operates across multiple Availability Zones. For disaster recovery, customers should implement automated service registration as part of their application startup process, ensuring services can re-register themselves when deployed in recovery regions. Custom scripts or infrastructure-as-code templates can be used to recreate namespace and service configurations.

Backup Capabilities

AWS Cloud Map does not have traditional backup capabilities or integration with AWS Backup since it is a dynamic service discovery registry. The service data (namespaces,

services, and instances) is designed to be transient and automatically managed by applications:

- Instead of backups, customers should implement infrastructure-as-code using AWS CloudFormation or other tools to recreate Cloud Map configurations
- Service instance data is expected to be re-registered automatically by applications during startup, making traditional backup mechanisms unnecessary for this service type

Disaster Recovery

AWS Cloud Map does not directly integrate with AWS Elastic Disaster Recovery as it is not a traditional data service requiring replication. However, the service itself provides high availability by running across multiple Availability Zones within each region:

- For disaster recovery scenarios, customers should deploy Cloud Map resources in multiple regions and implement application logic to re-register services automatically in the recovery region
- The service configurations can be replicated across regions using infrastructure-as-code tools, while service instances should be dynamically registered by applications as part of the disaster recovery process

Recovery Objectives

AWS Cloud Map helps customers meet RPO and RTO objectives by providing rapid service discovery and registration capabilities during disaster recovery scenarios. Since service instance data is ephemeral and designed for dynamic registration, the RPO is effectively zero as no historical data needs to be recovered. The RTO is primarily determined by how quickly applications can re-register services in the recovery region and discover each other, typically within minutes. The service's high availability architecture and fast DNS propagation help minimize discovery-related delays during failover events, supporting aggressive RTO targets for microservices architectures.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/cloud-map/latest/dg/cloud-map-limits.html

FAQ: https://aws.amazon.com/cloud-map/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/cloud-map/latest/dg/what-is-cloud-map.html

User Guide: https://docs.aws.amazon.com/cloud-map/latest/dg/tutorials.html

API Reference: https://docs.aws.amazon.com/cloud-map/latest/api/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/cloud-map/pricing/

Cost Optimization

AWS Cloud Map offers several cost optimization opportunities through its flexible discovery mechanisms and efficient resource management. Customers can reduce costs by choosing HTTP namespaces for API-only discovery, avoiding DNS-related charges when DNS discovery is not needed:

- Using appropriate DNS TTL values balances discovery freshness with query costs
- The service charges only for registered resources and API calls, so deregistering unused instances immediately reduces costs
- Implementing efficient service discovery patterns with caching and connection pooling minimizes API call frequency
- Cross-account namespace sharing eliminates duplicate registry costs in multi-account architectures, while service-level attributes reduce the need for additional discovery calls by providing metadata in single API responses

Savings Considerations

Cost savings with AWS Cloud Map come from eliminating the need to build and maintain custom service discovery infrastructure, reducing operational overhead and development time. The pay-per-use model means customers only pay for active registered resources and actual discovery calls, avoiding fixed infrastructure costs:

- Compared to running self-managed service registries on EC2 instances, Cloud Map eliminates compute, storage, and maintenance costs while providing higher availability
- The integration with container services reduces manual service management overhead
- DNS-based discovery can be more cost-effective than API calls for high-frequency lookups, while the free tier provides cost savings for smaller workloads with up to certain usage limits

Service Name

AWS CodeBuild

Service Overview

AWS CodeBuild is a fully managed build service in the cloud that compiles source code, runs unit tests, and produces artifacts ready for deployment. It eliminates the need to provision, manage, and scale build servers while providing prepackaged build environments for popular programming languages and build tools such as Apache Maven, Gradle, and more. CodeBuild automatically scales to meet peak build demands and charges based on build minutes consumed. The service integrates seamlessly with AWS CodePipeline for continuous delivery workflows and supports various source providers including GitHub, Bitbucket, AWS CodeCommit, and Amazon S3. It offers preconfigured environments for Microsoft Windows and Linux, with the flexibility to bring custom build environments using Docker containers.

Key Use Cases

- Continuous Integration: Automatically compile source code, run tests, and validate code quality as part of CI/CD pipelines
- Automated Testing: Execute unit tests, functional tests, and integration tests with visual reporting of test case results
- Multi-Environment Builds: Build applications across different programming languages and platforms using preconfigured or custom Docker environments
- Parallel Processing: Run test suites concurrently using parallel test execution to significantly reduce build times
- Container Image Building: Build and publish Docker images to Amazon ECR with dedicated Docker server capabilities and persistent caching

Features

- **Fully Managed Build Service**: Eliminates need to set up, patch, update, and manage build servers with automatic scaling
- **Multiple Source Providers**: Supports GitHub, GitHub Enterprise, Bitbucket, AWS CodeCommit, and Amazon S3 as source providers
- **Preconfigured Build Environments**: Provides ready-to-use environments for popular programming languages and build tools
- **Custom Docker Environments**: Allows bringing customized build environments as Docker containers
- **Parallel Test Execution**: Enables running test suites concurrently to reduce build times significantly

- **Test Reporting**: Generates comprehensive reports for unit, functional, and integration tests with visual results
- **VPC Integration**: Can operate within Amazon VPC for integration with private services and databases
- **Docker Server Capability**: Provides dedicated Docker server with persistent cache for faster container builds
- **Build Caching**: Supports caching of build dependencies to improve performance across builds

Value Proposition

AWS CodeBuild provides significant advantages over self-managed build infrastructure by eliminating server management overhead and offering automatic scaling without capacity planning. The pay-per-use model ensures cost efficiency as customers only pay for build minutes consumed rather than maintaining idle infrastructure:

- CodeBuild's seamless integration with AWS developer tools creates comprehensive CI/CD pipelines, while preconfigured environments for popular languages reduce setup time
- Advanced features like parallel test execution, Docker server capabilities, and persistent caching dramatically improve build performance
- The service's ability to operate within VPCs enables secure integration with existing infrastructure, making it ideal for enterprises requiring both flexibility and security in their build processes

Service Integration

AWS CodeBuild integrates deeply with AWS CodePipeline as a build and test stage in continuous delivery workflows, enabling automated code release processes. It works seamlessly with AWS CodeCommit for source control and AWS CodeArtifact for artifact management:

- Integration with Amazon S3 provides artifact storage and source input capabilities, while AWS Identity and Access Management (IAM) controls access to builds and projects
- Amazon CloudWatch and CloudWatch Logs provide comprehensive monitoring and logging of build activities
- CodeBuild also integrates with Amazon SNS for build notifications, AWS Systems Manager Parameter Store and AWS Secrets Manager for secure environment variable storage, and can operate within Amazon VPC for private resource access
- Additional integrations include Amazon ECR for container image storage and AWS Lambda for serverless build operations

Onboarding and Offboarding

Getting started with AWS CodeBuild involves creating a build project that defines source code location, build environment, and output settings. Customers begin by planning their build requirements, including determining source repository providers, build commands, required runtimes and tools, and any additional AWS resources needed:

- The service supports source code from CodeCommit, Amazon S3, GitHub, and Bitbucket
- Users create a buildspec file in YAML format that instructs CodeBuild on build transformation processes
- Build projects can be created through the AWS console, CLI, or SDKs, with options to configure environment settings, artifact outputs, and logging
- CodeBuild provides step-by-step tutorials for common scenarios, including Java applications with Maven, demonstrating the complete build lifecycle from source upload to artifact retrieval

Getting Started Guide: https://docs.aws.amazon.com/codebuild/latest/userguide/getting-started-overview.html

Data Removal

When offboarding from AWS CodeBuild, customers can delete build projects, builds, and associated resources using the AWS console, CLI, or SDKs. Build projects with existing builds and resource policies require removal of the policy and deletion of builds before project deletion. The batch-delete-builds API allows deletion of multiple builds simultaneously. All build artifacts stored in S3 buckets, CloudWatch logs, and test reports are retained according to their respective service retention policies and must be managed separately. Complete data removal requires deleting all associated resources including IAM roles, S3 artifacts, and CloudWatch log groups created for CodeBuild operations.

Backup/Restore and Disaster Recovery

AWS CodeBuild inherently provides resilience through its serverless, fully managed architecture that operates across multiple Availability Zones. Build configurations and project definitions are automatically replicated across AWS infrastructure. For disaster recovery, customers should implement Infrastructure as Code using AWS CloudFormation or CDK to recreate CodeBuild projects and configurations. Source code repositories and build artifacts should be backed up separately through their respective services. Build history and logs are retained according to CloudWatch retention policies, and customers can configure cross-region replication for critical build artifacts stored in S3 buckets.

Backup Capabilities

CodeBuild does not directly integrate with AWS Backup service as it is a serverless, fully managed service. However, the service maintains build project configurations and metadata through AWS's standard infrastructure resilience:

- Customers are responsible for backing up source code repositories, build artifacts in S3, and any custom Docker images in ECR
- Build specifications and project configurations should be version-controlled and backed up through Infrastructure as Code practices
- CloudWatch logs and build history are retained according to configured retention periods, providing historical build information for recovery purposes

Disaster Recovery

CodeBuild does not directly integrate with AWS Elastic Disaster Recovery as it is a serverless service with inherent multi-AZ resilience. The service automatically handles infrastructure failures and provides high availability through AWS's managed infrastructure:

- For comprehensive disaster recovery, customers should implement cross-region strategies by replicating build projects using CloudFormation templates, maintaining source code repositories in multiple regions, and storing critical build artifacts in geographically distributed S3 buckets
- Build configurations can be quickly recreated in alternate regions using Infrastructure as Code approaches, ensuring business continuity during regional outages

Recovery Objectives

CodeBuild supports customer RPO and RTO objectives through its serverless architecture that eliminates infrastructure recovery time. Build projects and configurations can be instantly recreated from Infrastructure as Code templates, achieving near-zero RTO for project recreation. Source code and artifacts should be replicated according to business requirements to meet RPO objectives. The service's integration with multi-region source repositories and S3 cross-region replication enables customers to maintain low RPO targets. For critical builds, customers can implement automated cross-region project creation and maintain standby configurations to minimize recovery time during disaster scenarios.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/codebuild/latest/userguide/limits.html

FAQ: https://aws.amazon.com/codebuild/faqs/

Technical Requirements

Service Documentation:
https://docs.aws.amazon.com/codebuild/latest/userguide/welcome.html

User Guide: https://docs.aws.amazon.com/codebuild/latest/userguide/welcome.html

API Reference:
https://docs.aws.amazon.com/codebuild/latest/APIReference/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/codebuild/pricing/

Cost Optimization

AWS CodeBuild offers several unique cost optimization opportunities through its pay-per-use pricing model where customers only pay for build minutes consumed. Reserved capacity instances provide cost savings for predictable workloads with minimum usage commitments:

- The service's auto-scaling eliminates costs of idle infrastructure common with self-managed build servers
- Parallel test execution reduces total build time, directly decreasing costs
- Docker server capability with persistent caching significantly reduces build times for container-based workflows
- Customers can optimize costs by selecting appropriate compute types, using efficient build scripts, implementing build caching strategies, and leveraging the free tier offering 100 build minutes monthly for small instances
- Strategic use of spot pricing for non-critical builds and proper resource right-sizing further optimize expenses

Savings Considerations

Primary cost savings with CodeBuild come from eliminating the need to provision, manage, and maintain dedicated build infrastructure, reducing both capital and operational expenses. The pay-per-use model ensures customers avoid costs of idle build servers, paying only for actual build time consumed:

- Automatic scaling prevents over-provisioning while meeting peak demands efficiently
- Advanced features like parallel test execution and Docker server caching dramatically reduce build durations, directly translating to lower costs

- The service eliminates server maintenance, patching, and update costs while providing enterprise-grade build capabilities
- Integration with other AWS services creates operational efficiencies that reduce overall development lifecycle costs
- Organizations typically see 30-60% cost reduction compared to self-managed build infrastructure while gaining improved performance and reliability

Service Name

AWS CodePipeline

Service Overview

AWS CodePipeline is a fully managed continuous delivery service that automates and models the software release process. It enables users to visualize, configure, and automate the different stages of a software release workflow, from source code changes to production deployment. CodePipeline detects changes in source repositories, automatically builds code, runs tests, and deploys applications to production environments. The service orchestrates the entire release process through a series of configurable stages and actions, providing a consistent and reliable way to release software changes rapidly with high frequency and quality.

Key Use Cases

- Continuous Integration/Continuous Deployment (CI/CD): Automate build, test, and deployment workflows for applications across development, staging, and production environments
- Multi-stage application deployment: Deploy applications to Amazon EC2 instances, AWS Elastic Beanstalk, Amazon ECS containers, and AWS Lambda functions with automated progression through environments
- Infrastructure as Code automation: Automate deployment and management of AWS resources using AWS CloudFormation templates with version control and rollback capabilities

Features

- **Visual Pipeline Modeling**: Model release workflows using visual pipelines with multiple configurable stages and actions
- **AWS Service Integrations**: Native integration with AWS CodeCommit, CodeBuild, CodeDeploy, Elastic Beanstalk, ECS, Lambda, and CloudFormation
- **Stage-level Conditions**: Configure entry, success, and failure conditions with automatic rollback capabilities for enhanced release control
- **Pipeline Variables**: Use pipeline-level and action-level variables for dynamic configuration and parameterized deployments
- **Parallel and Sequential Execution**: Execute actions in parallel within stages or sequentially across stages with flexible execution modes
- **Cross-Region Support**: Deploy applications across multiple AWS regions using cross-region actions and artifact replication
- **Third-party Integrations**: Integration with GitHub, Jenkins, and other third-party tools through prebuilt and custom plugins

- **Manual Approval Gates**: Include manual approval actions to control progression between pipeline stages

Value Proposition

CodePipeline provides a fully managed, scalable CI/CD solution that eliminates the need to maintain custom deployment infrastructure. It offers deep integration with the AWS ecosystem, enabling seamless automation from source to production with minimal operational overhead:

- The service provides enterprise-grade features like stage-level conditions, automatic rollbacks, and cross-account deployments while maintaining cost-effectiveness through its pay-per-use pricing model
- Unlike self-managed CI/CD solutions, CodePipeline automatically scales, provides high availability, and integrates natively with AWS security and compliance features, reducing time-to-market and operational complexity for development teams

Service Integration

CodePipeline integrates extensively with core AWS developer services including AWS CodeCommit for source control, AWS CodeBuild for build and test automation, and AWS CodeDeploy for application deployment. It connects with compute services like Amazon EC2, AWS Lambda, Amazon ECS, and AWS Elastic Beanstalk for deployment targets:

- The service integrates with AWS CloudFormation for infrastructure-as-code deployments and Amazon S3 for artifact storage
- Additional integrations include Amazon SNS for notifications, AWS IAM for security and access control, Amazon CloudWatch for monitoring and logging, and AWS CloudTrail for API auditing and compliance tracking

Onboarding and Offboarding

Customers can get started by creating an AWS account and setting up an administrative user with CodePipeline permissions. The service offers a simplified getting started experience with pre-built pipeline templates for Build, Automation, and Deployment use cases:

- Users can create their first pipeline using the AWS Management Console, selecting a source repository (CodeCommit, GitHub, or S3), adding build and deployment stages, and configuring the target environment
- The new pipeline templates allow quick creation of fully configured pipelines ready to run
- Tutorials are available for common scenarios like deploying to EC2 instances, ECS containers, and Lambda functions

Getting Started Guide:
https://docs.aws.amazon.com/codepipeline/latest/userguide/getting-started-codepipeline.html

Data Removal

To offboard from CodePipeline, customers must delete all pipelines, which automatically removes pipeline execution history and associated metadata. Pipeline artifacts stored in Amazon S3 buckets must be manually deleted as they are retained independently. Custom actions, webhooks, and connections should be removed through the console or API. IAM roles and policies created for CodePipeline can be deleted once no longer needed. The service retains pipeline execution history for 12 months before automatic deletion, so customers should export any required historical data before offboarding.

Backup/Restore and Disaster Recovery

CodePipeline does not provide traditional backup capabilities as it is a workflow orchestration service that manages deployment processes rather than persistent data. Pipeline configurations can be exported and recreated using AWS CloudFormation templates or the AWS CLI for disaster recovery purposes. The service automatically maintains pipeline execution history for 12 months. For disaster recovery, customers should replicate pipeline configurations across regions and ensure source repositories and deployment targets have appropriate backup and recovery mechanisms in place.

Backup Capabilities

CodePipeline itself does not have backup capabilities and does not integrate with AWS Backup as it manages workflow orchestration rather than persistent data storage. Pipeline configurations serve as the primary ‘data’ and can be version-controlled through Infrastructure as Code practices using CloudFormation or CDK. Customers should backup pipeline definitions, source code repositories, and deployment artifacts separately using appropriate backup services for each component.

Disaster Recovery

CodePipeline does not directly integrate with AWS Elastic Disaster Recovery as it orchestrates deployment workflows rather than managing application infrastructure. However, pipelines can be configured to deploy across multiple regions and support disaster recovery strategies for the applications they deploy. Customers should implement cross-region pipeline configurations and ensure that source repositories, build artifacts, and deployment targets have appropriate disaster recovery mechanisms configured independently.

Recovery Objectives

CodePipeline helps achieve faster Recovery Time Objectives (RTO) by automating deployment processes and enabling rapid application updates across multiple

environments. The service supports automated rollback capabilities through stage-level conditions, reducing Recovery Point Objectives (RPO) by quickly reverting failed deployments. Cross-region deployment capabilities enable geographic distribution for improved disaster recovery posture. However, specific RTO/RPO targets depend on the underlying services and applications being deployed rather than CodePipeline itself.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/general/latest/gr/codepipeline.html

FAQ: https://aws.amazon.com/codepipeline/faqs/

Technical Requirements

Service Documentation:
https://docs.aws.amazon.com/codepipeline/latest/userguide/welcome.html

User Guide: https://docs.aws.amazon.com/codepipeline/latest/userguide/welcome.html

API Reference:
https://docs.aws.amazon.com/codepipeline/latest/APIReference/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/codepipeline/pricing/

Cost Optimization

CodePipeline offers two pricing models: V1 pipelines at $1.00 per active pipeline per month (flat rate) and V2 pipelines at $0.002 per action execution minute (usage-based). Cost optimization strategies include choosing the appropriate pipeline type based on usage patterns, utilizing the AWS Free Tier (one free V1 pipeline monthly, 100 free V2 action execution minutes), and minimizing unnecessary pipeline executions through efficient triggering mechanisms. Organizations can reduce costs by consolidating similar workflows, optimizing build times, and using efficient artifact storage practices in S3.

Savings Considerations

Primary cost savings come from eliminating the need to maintain self-managed CI/CD infrastructure, reducing operational overhead and staffing requirements. The pay-per-use model for V2 pipelines provides cost efficiency for variable workloads compared to maintaining always-on build servers:

- Organizations save on licensing costs for third-party CI/CD tools and reduce time-to-market through automated deployments

- Additional savings result from improved deployment reliability, reduced rollback costs, and decreased manual intervention requirements
- The service's integration with other AWS services can provide economies of scale and volume discounts across the broader AWS usage portfolio

Service Name

AWS Compute Optimizer

Service Overview

AWS Compute Optimizer is a machine learning-powered service that analyzes the configuration and utilization metrics of AWS resources to provide optimization recommendations. It evaluates Amazon EC2 instances, Auto Scaling groups, EBS volumes, AWS Lambda functions, Amazon ECS services on Fargate, commercial software licenses, and Amazon Aurora/RDS databases to identify right-sizing opportunities, idle resources, and cost optimization potential. The service analyzes historical utilization patterns and provides actionable recommendations to reduce costs and improve performance by suggesting optimal resource configurations based on actual usage data.

Key Use Cases

- Cost optimization: Identify over-provisioned and idle resources to reduce infrastructure costs by rightsizing or eliminating unnecessary resources
- Performance optimization: Recommend instance types and configurations that better match workload requirements to improve application performance
- Resource rightsizing: Analyze utilization patterns across EC2 instances, Auto Scaling groups, Lambda functions, EBS volumes, and databases to suggest optimal sizing
- License optimization: Optimize commercial software licenses like Microsoft SQL Server on EC2 to reduce licensing costs
- Multi-account optimization: Centrally manage optimization recommendations across AWS Organizations with delegated administration capabilities

Features

- **Resource Analysis**: Analyzes EC2 instances, Auto Scaling groups, EBS volumes, Lambda functions, ECS services, and RDS/Aurora databases using machine learning and utilization metrics
- **Automation Rules**: Creates automation rules to automatically implement optimization recommendations based on criteria like AWS Region and resource tags on recurring schedules
- **Enhanced Infrastructure Metrics**: Extends analysis lookback period from 14 days to 93 days for more accurate quarterly utilization pattern detection
- **Idle Resource Detection**: Identifies unused or underutilized resources like unattached EBS volumes and idle EC2 instances for cleanup
- **Performance Risk Assessment**: Evaluates performance risk of recommended changes with ratings from very low to very high based on utilization patterns

- **Savings Estimation**: Calculates potential cost savings and considers pricing discounts when generating recommendations for cost optimization
- **External Metrics Integration**: Ingests metrics from external observability products to enhance recommendation accuracy
- **Organization Integration**: Integrates with AWS Organizations for centralized management and delegated administration across multiple accounts

Value Proposition

AWS Compute Optimizer provides machine learning-powered optimization recommendations that deliver measurable cost savings and performance improvements without requiring extensive manual analysis. Unlike basic monitoring tools, it offers automated rightsizing recommendations with performance risk assessment, ensuring changes won't negatively impact applications:

- The service provides automation capabilities to implement recommendations at scale across organizations, extended analytics with 93-day lookback periods for seasonal patterns, and integration with existing AWS services
- It's particularly valuable because it considers actual utilization patterns rather than theoretical capacity, offers reversible automated actions for operational safety, and provides detailed savings estimates with support for custom pricing discounts

Service Integration

AWS Compute Optimizer integrates deeply with Amazon CloudWatch for metrics collection and analysis, AWS Organizations for multi-account management, and AWS Cost Explorer for unified cost optimization views. It works with Amazon EC2, Auto Scaling groups, Lambda, ECS on Fargate, EBS, Aurora, and RDS for resource optimization:

- The service integrates with AWS Trusted Advisor for complementary recommendations, AWS Systems Manager for automation workflows, and AWS Application Discovery Service for workload assessment
- It also connects with AWS Backup for data protection during optimization, AWS Resource Groups for organized management, and external observability platforms for enhanced metrics ingestion

Onboarding and Offboarding

Customers get started by opting into AWS Compute Optimizer through the console, CLI, or API, which automatically creates a service-linked role for data access. The service supports standalone accounts, organization member accounts, and management accounts with options to include all organization members:

- After opting in, Compute Optimizer begins analyzing resources with sufficient metrics data (30 hours for EC2/EBS, 50 invocations for Lambda)

- Enhanced infrastructure metrics can be activated for deeper 93-day analysis
- Organizations can designate delegated administrators and configure automation rules for scaled optimization
- The service requires appropriate IAM permissions and supports trusted access for organization-wide management

Getting Started Guide: https://docs.aws.amazon.com/compute-optimizer/latest/ug/getting-started.html

Data Removal

Customers can opt out of AWS Compute Optimizer through the console, CLI, or API, which stops new analysis and recommendation generation. To remove historical data, customers must contact AWS Support to request deletion of stored CloudWatch metrics and configuration data used for recommendations. The service automatically stops accessing resources upon opt-out, and service-linked roles can be deleted if no longer needed.

Backup/Restore and Disaster Recovery

AWS Compute Optimizer does not provide traditional backup and disaster recovery capabilities as it is an optimization service rather than a data storage service. The service itself is highly available across AWS regions and maintains recommendation data, but does not back up customer resources. However, automation features include safety mechanisms like creating EBS snapshots before volume modifications to enable data recovery if needed.

Backup Capabilities

AWS Compute Optimizer does not have inherent backup capabilities and does not integrate with AWS Backup as it is an analysis and recommendation service rather than a data service. The service focuses on optimization recommendations rather than data protection, though automation features include safety mechanisms like snapshot creation before resource modifications.

Disaster Recovery

AWS Compute Optimizer does not directly support disaster recovery scenarios and does not integrate with AWS Elastic Disaster Recovery. As an optimization service, it focuses on cost and performance recommendations rather than business continuity. The service itself is regionally distributed for high availability but does not provide disaster recovery capabilities for customer workloads.

Recovery Objectives

AWS Compute Optimizer does not directly help customers meet RPO and RTO objectives as it is not a backup or disaster recovery service. However, it can indirectly support resilience by recommending properly sized resources for better performance and

identifying idle resources that could be repurposed for disaster recovery scenarios in multi-region architectures.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/general/latest/gr/compute-optimizer.html

FAQ: https://aws.amazon.com/compute-optimizer/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/compute-optimizer/latest/ug/what-is-compute-optimizer.html

User Guide: https://docs.aws.amazon.com/compute-optimizer/latest/ug/getting-started.html

API Reference: https://docs.aws.amazon.com/compute-optimizer/latest/APIReference/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/compute-optimizer/pricing/

Cost Optimization

AWS Compute Optimizer offers unique cost optimization through machine learning-driven rightsizing recommendations that consider actual utilization patterns rather than theoretical capacity. The service provides automated cleanup of idle resources like unattached EBS volumes, license optimization for commercial software, and personalized savings estimation that accounts for customer-specific pricing discounts and Reserved Instances. Enhanced infrastructure metrics enable quarterly pattern analysis for seasonal workloads, while automation rules allow scaled optimization across entire organizations with reversible actions for operational safety and confidence in implementation.

Savings Considerations

Primary cost savings come from rightsizing over-provisioned resources, eliminating idle resources, and optimizing commercial software licenses. The service identifies instances that can be downsized while maintaining performance, recommends modern instance types with better price-performance ratios, and suggests migration to AWS Graviton instances for additional savings:

- Enhanced metrics capture seasonal patterns to prevent under-provisioning during peak periods

- Automation features enable continuous optimization at scale, while performance risk assessment ensures cost reductions don’t compromise application performance, maximizing sustainable long-term savings opportunities

Service Name

AWS Config

Service Overview

AWS Config is a fully managed service that provides a detailed view of AWS resource configurations in an account, including how resources are related to one another and their historical configurations. It enables organizations to assess, audit, and evaluate the configurations of AWS resources by continuously monitoring and recording configuration changes as configuration items. AWS Config helps maintain visibility into resource relationships, tracks configuration drift, ensures compliance through rules and conformance packs, and provides centralized governance across multiple accounts and regions through aggregators.

Key Use Cases

- Compliance monitoring: Continuously monitor resource configurations against industry standards, security best practices, and organizational policies using AWS Config rules and conformance packs
- Configuration management: Track and audit configuration changes across AWS resources to maintain visibility, troubleshoot issues, and understand resource relationships
- Security governance: Assess security postures by evaluating resource configurations against desired states, identifying vulnerabilities, and enabling automated remediation of non-compliant resources

Features

- **Configuration Recording**: Records configuration changes for supported AWS resources as configuration items with point-in-time snapshots
- **AWS Config Rules**: Evaluates resource compliance against desired configurations using managed or custom rules with automatic evaluation
- **Conformance Packs**: Bundles multiple rules and remediation actions for compliance against industry standards and best practices
- **Multi-Account Aggregation**: Aggregates configuration and compliance data across multiple AWS accounts, regions, and organizations
- **Advanced Query**: Provides SQL-like queries to search and analyze current configuration state across resources
- **Automated Remediation**: Automatically fixes non-compliant resources using AWS Systems Manager documents or Lambda functions
- **Resource Relationships**: Maps and tracks relationships between AWS resources to understand dependencies and impact

Value Proposition

AWS Config provides unique value through its comprehensive configuration management and compliance automation capabilities that are deeply integrated with the AWS ecosystem. Unlike third-party solutions, Config offers native integration with over 150 AWS services, enabling real-time configuration tracking without performance impact:

- Its managed rules library covers industry standards like PCI DSS, HIPAA, and CIS benchmarks, reducing compliance overhead
- The service scales automatically across multi-account environments with centralized aggregation, provides sub-second configuration change detection, and integrates seamlessly with AWS Security Hub, CloudTrail, and Systems Manager for holistic governance

Service Integration

AWS Config integrates deeply with core AWS services to provide comprehensive governance capabilities. It works with AWS Organizations for multi-account management, AWS Control Tower for automated compliance in landing zones, and AWS CloudTrail to correlate configuration changes with API events:

- Security integrations include AWS Security Hub for centralized security findings, AWS Audit Manager for evidence collection, and AWS Systems Manager for automated remediation
- Config also integrates with Amazon S3 for configuration storage, Amazon SNS for notifications, CloudWatch for monitoring, AWS Firewall Manager for security policy compliance, and AWS Trusted Advisor for operational recommendations

Onboarding and Offboarding

Customers can start with AWS Config using the 1-click setup in the AWS Management Console, which automatically configures best-practice settings for resource recording, delivery channels, and IAM roles. The setup process includes selecting recording strategies (continuous or daily), choosing resource types to monitor, configuring S3 buckets for configuration storage, and optionally setting up SNS topics for notifications:

- For multi-account environments, customers can deploy Config using AWS CloudFormation, Terraform, or AWS Systems Manager Quick Setup
- AWS Control Tower automatically enables Config in enrolled accounts across all regions

Getting Started Guide:
https://docs.aws.amazon.com/config/latest/developerguide/getting-started.html

Data Removal

AWS Config data can be removed by configuring retention periods between 30 days and 7 years for configuration items, with a default of 7 years. Configuration recorders can be stopped and deleted to cease new data collection. Historical configuration data stored in customer-owned S3 buckets must be manually deleted. When offboarding, customers should delete Config rules, conformance packs, aggregators, and delivery channels. The service automatically stops generating new configuration items when recorders are stopped.

Backup/Restore and Disaster Recovery

AWS Config stores configuration data in customer-owned Amazon S3 buckets, providing durability and availability through S3's built-in redundancy. Configuration items are automatically backed up with customizable retention periods up to 7 years. The service supports cross-region replication of configuration data through S3 bucket replication. However, AWS Config primarily focuses on configuration tracking rather than operational backup and disaster recovery capabilities for customer workloads.

Backup Capabilities

AWS Config does not provide traditional backup capabilities but maintains historical configuration data through configuration items stored in S3 buckets. The service retains configuration history for up to 7 years and does not integrate directly with AWS Backup service. Configuration snapshots can be generated on-demand and stored in S3 for point-in-time recovery of configuration states, but this is for configuration data only, not application or user data.

Disaster Recovery

AWS Config does not directly integrate with AWS Elastic Disaster Recovery as it focuses on configuration management rather than operational disaster recovery. The service achieves high availability through AWS's global infrastructure and stores configuration data in durable S3 buckets. For disaster recovery scenarios, customers can replicate Config aggregators and rules across regions, but the service is primarily designed for governance and compliance rather than operational recovery.

Recovery Objectives

AWS Config helps customers meet Recovery Point Objectives (RPO) by maintaining detailed configuration history up to 7 years, enabling point-in-time configuration recovery. For Recovery Time Objectives (RTO), Config provides rapid configuration state queries through Advanced Query API and aggregators for quick assessment during incidents. The service supports automated compliance evaluation and remediation, helping reduce recovery time by ensuring consistent configurations that meet operational requirements.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/config/latest/developerguide/configlimits.html

FAQ: https://aws.amazon.com/config/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/config/latest/developerguide/WhatIsConfig.html

User Guide: https://docs.aws.amazon.com/config/latest/developerguide/getting-started.html

API Reference: https://docs.aws.amazon.com/config/latest/APIReference/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/config/pricing/

Cost Optimization

AWS Config offers several cost optimization strategies including selective resource recording to monitor only critical resources, periodic recording (daily vs. continuous) for less critical resources, and configurable retention periods from 30 days to 7 years to balance compliance needs with storage costs:

- Customers can optimize costs by excluding non-essential resource types from recording, using frequency overrides for different resource categories, and leveraging managed rules instead of custom Lambda-based rules to reduce compute costs
- The service integrates with Cost Explorer for tracking Config-related expenses

Savings Considerations

Cost savings with AWS Config can be achieved through strategic configuration recording approaches. Customers should implement periodic recording for static resources and continuous recording only for critical, frequently-changing resources:

- Using managed rules reduces Lambda execution costs compared to custom rules
- Proper retention period configuration prevents unnecessary long-term storage costs
- Organizations can achieve economies of scale through multi-account aggregation, reducing redundant rule evaluations across accounts

- Proactive compliance monitoring prevents costly misconfigurations and security incidents that could result in significant operational expenses

Service Name

AWS Control Tower

Service Overview

AWS Control Tower offers a straightforward way to set up and govern an AWS multi-account environment following prescriptive best practices. It orchestrates capabilities of several AWS services including AWS Organizations, AWS Service Catalog, and AWS IAM Identity Center to build a secure, compliant landing zone in less than an hour. The service helps organizations maintain compliance and security standards across multiple accounts while providing account factory capabilities for rapid account provisioning. It implements preventive, detective, and proactive controls to ensure ongoing governance and prevent drift from best practices.

Key Use Cases

- Multi-account environment setup and governance for enterprises requiring standardized security and compliance controls
- Account Factory provisioning to enable distributed teams to quickly create new AWS accounts using configurable templates
- Regulatory compliance management through automated implementation of security guardrails and policy enforcement across organizational units

Features

- **Landing Zone**: Well-architected multi-account environment based on security and compliance best practices that scales to enterprise needs
- **Controls (Guardrails)**: High-level rules providing ongoing governance with preventive, detective, and proactive control types that can be mandatory, strongly recommended, or elective
- **Account Factory**: Configurable account template that standardizes provisioning of new accounts with pre-approved configurations and automated workflows
- **Dashboard**: Provides continuous oversight allowing administrators to see provisioned accounts, enabled controls, and noncompliant resources organized by accounts and OUs
- **Control Catalog**: Over 750 managed controls with support for various compliance frameworks and configurable exemptions for specific resources or IAM principals
- **Automated Account Enrollment**: Automatically enrolls accounts and applies appropriate baselines and controls when accounts are moved between organizational units

Value Proposition

AWS Control Tower provides the easiest way to set up and govern a secure, compliant multi-account AWS environment based on best practices established through thousands of enterprise engagements. It offers automated governance at scale, eliminating manual configuration overhead while ensuring consistent security posture:

- The service provides built-in Account Factory for rapid account provisioning, comprehensive control catalog with over 750 managed controls, and seamless integration with existing AWS Organizations structures
- It reduces operational complexity through centralized management while maintaining flexibility for customization and extensibility across different organizational requirements

Service Integration

AWS Control Tower deeply integrates with AWS Organizations for multi-account management and centralized billing, AWS Service Catalog for Account Factory provisioning workflows, and AWS IAM Identity Center for single sign-on capabilities. Core integrations include AWS Config for configuration governance and compliance monitoring, AWS CloudTrail for activity logging and audit trails, AWS CloudFormation StackSets for resource deployment across accounts, and AWS Security Hub for centralized security findings. Additional integrations encompass AWS Backup for data protection, AWS Lambda for automation functions, Amazon S3 for logging storage, Amazon SNS for notifications, and AWS Step Functions for orchestrating complex workflows.

Onboarding and Offboarding

Customers begin by selecting their home region and completing automated pre-launch checks for their management account. Setup involves choosing between full landing zone deployment or controls-dedicated experience for existing multi-account environments:

- The quick start process creates audit and log archive accounts, establishes organizational units, and deploys initial controls in approximately 30 minutes
- Prerequisites include sufficient AWS service limits, active AWS support plan, and proper IAM permissions
- For existing organizations, AWS Control Tower creates parallel structures without disrupting current OUs and accounts, allowing immediate use alongside existing AWS Organizations environments

Getting Started Guide:
https://docs.aws.amazon.com/controltower/latest/userguide/getting-started-with-control-tower.html

Data Removal

Offboarding requires careful decommissioning of AWS Control Tower resources while preserving account structures and data. Process involves disabling enabled controls, unregistering organizational units, and removing landing zone components. Data retention follows standard AWS service policies for integrated services like CloudTrail, Config, and S3. Organizations must manually delete specific resources like S3 buckets containing logs and CloudFormation stacks to ensure complete removal.

Backup/Restore and Disaster Recovery

AWS Control Tower integrates with AWS Backup to provide automated, centralized backup capabilities across the landing zone environment. The service offers prescriptive backup plans with predefined retention policies, backup frequencies, and time windows applied to organizational units. Backup functionality requires setup of dedicated backup administrator and central backup accounts with multi-region KMS keys for encryption. Integration ensures consistent data protection standards across all governed member accounts.

Backup Capabilities

AWS Control Tower supports comprehensive backup integration through AWS Backup service with prescriptive backup plans at the landing zone level. Backup capabilities include automated resource tagging for backup inclusion, centralized backup vault creation, and four types of backup plans with configurable retention periods. The service enables consistent backup policies across organizational units with automatic inclusion for new accounts provisioned under registered OUs.

Disaster Recovery

Disaster recovery is supported through underlying AWS service capabilities rather than native Control Tower features. Organizations can implement backup and restore, pilot light, warm standby, or multi-site active/active strategies using integrated services. AWS Control Tower's multi-region governance capabilities support disaster recovery architectures by enabling consistent security controls and account structures across regions for resilience planning.

Recovery Objectives

AWS Control Tower assists with RPO and RTO objectives through its integrated backup capabilities and multi-region governance structure. Automated backup plans help achieve specific recovery point objectives, while consistent account structures and security controls across regions support faster recovery time objectives. The service's centralized governance model enables standardized disaster recovery implementations across the multi-account environment.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/controltower/latest/userguide/limits.html

FAQ: https://aws.amazon.com/controltower/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/controltower/latest/userguide/what-is-control-tower.html

User Guide: https://docs.aws.amazon.com/controltower/latest/userguide/getting-started-with-control-tower.html

API Reference: https://docs.aws.amazon.com/controltower/latest/APIReference/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/controltower/pricing/

Cost Optimization

AWS Control Tower itself incurs no additional charges as it's a free orchestration service. Cost optimization focuses on managing underlying AWS services usage including AWS Config configuration items, CloudTrail data events, and S3 storage for logs:

- Organizations can optimize costs by carefully selecting detective controls to avoid duplication, managing ephemeral workloads that generate excessive Config items, and leveraging organizational-level CloudTrail trails to prevent duplicate logging charges
- Proper tagging strategies enable accurate cost attribution across accounts and organizational units

Savings Considerations

Primary cost savings come from avoiding duplicate AWS service deployments through centralized governance and shared security infrastructure. Organizations benefit from consolidated billing across multiple accounts, optimized CloudTrail trail configurations at organization level, and standardized resource provisioning that prevents over-provisioning:

- The Account Factory reduces operational overhead costs by automating account creation and baseline application

- Centralized backup policies and security controls eliminate redundant implementations across accounts, while consistent governance reduces compliance audit costs and administrative overhead

Service Name

AWS Cost Explorer

Service Overview

AWS Cost Explorer is a comprehensive cost management and analysis tool that enables users to visualize, understand, and manage their AWS costs and usage over time. It provides interactive dashboards, customizable reports, and forecasting capabilities to help organizations analyze spending patterns, identify cost optimization opportunities, and make data-driven financial decisions. The service offers both a web-based user interface and programmatic access through APIs, supporting detailed cost analysis across multiple dimensions including services, accounts, regions, and time periods with up to 13 months of historical data and 18 months of forecasting.

Key Use Cases

- Cost Analysis and Monitoring: Track and analyze AWS spending patterns across services, accounts, and regions to identify cost drivers and trends
- Budget Planning and Forecasting: Generate cost forecasts for up to 18 months and create budgets with automated alerts for proactive financial management
- Cost Optimization: Identify underutilized resources, receive Reserved Instance recommendations, and analyze cost anomalies to reduce unnecessary spending

Features

- **Interactive Cost Visualization**: Customizable charts and graphs for exploring cost and usage data with filtering and grouping options
- **18-Month Forecasting**: AI-powered cost forecasting with explainable insights based on historical usage patterns
- **Cost Comparison**: Automated cost comparison between time periods with detailed breakdown of cost drivers
- **Granular Data Access**: Daily, hourly, and resource-level cost data with support for multi-year historical analysis
- **Reserved Instance Reports**: Coverage and utilization reporting with purchase recommendations for Reserved Instances and Savings Plans
- **Custom Reports and Dashboards**: Saveable custom reports and integration with Billing and Cost Management Dashboards
- **API Access**: Programmatic access to cost and usage data through comprehensive REST APIs
- **CSV Export**: Downloadable reports in CSV format for external analysis and integration

Value Proposition

AWS Cost Explorer offers unmatched visibility into AWS spending with native integration across all AWS services and real-time data access. Unlike third-party tools, it provides built-in forecasting with AI-powered explanations, automated cost anomaly detection, and seamless integration with AWS Budgets and Cost Anomaly Detection:

- The service eliminates the need for external cost management infrastructure while offering both free console access and affordable API pricing at $0.01 per request
- Its comprehensive feature set includes 18-month forecasting, resource-level granularity, and automated Reserved Instance recommendations, making it the most cost-effective solution for AWS-centric organizations seeking complete financial visibility and control

Service Integration

Cost Explorer integrates deeply with AWS Billing and Cost Management services including AWS Budgets for alert management, AWS Cost Anomaly Detection for automated spending anomaly identification, and AWS Cost and Usage Reports for detailed billing data. It connects with AWS Organizations for consolidated billing analysis across member accounts and integrates with Amazon Q Developer for natural language cost queries:

- The service also works with AWS IAM for access control, Amazon S3 for data export, and AWS Trusted Advisor for cost optimization recommendations
- Additionally, it supports integration with Amazon QuickSight for advanced analytics and AWS Config for resource tracking and compliance monitoring

Onboarding and Offboarding

Getting started with Cost Explorer is automatic upon AWS account creation, requiring no additional setup or configuration. Users simply access the service through the AWS Cost Management console where data preparation begins immediately, with current month data available within 24 hours and historical data within a few days:

- The service provides preconfigured default reports and an interactive walkthrough for new users
- Cost Explorer cannot be disabled once enabled, ensuring continuous cost visibility
- Organizations using consolidated billing need management account enablement for member account access, and IAM permissions can be configured for granular access control across teams and departments

Getting Started Guide: https://docs.aws.amazon.com/cost-management/latest/userguide/ce-getting-started.html

Data Removal

Cost Explorer cannot be disabled once enabled and maintains cost data according to AWS data retention policies. Historical cost data is retained for up to 13 months automatically, with optional extended retention for multi-year analysis. Since Cost Explorer uses the same dataset as AWS Cost and Usage Reports, data removal follows standard AWS account closure procedures. Organizations can control access through IAM policies and remove saved reports and custom configurations, but the underlying cost data remains accessible as part of AWS's billing infrastructure.

Backup/Restore and Disaster Recovery

Cost Explorer does not require traditional backup and disaster recovery as it operates as a fully managed AWS service with built-in redundancy and high availability. The service automatically maintains data integrity and availability across AWS regions. Cost data is continuously synchronized from billing systems and refreshed at least once every 24 hours. Custom reports and configurations can be recreated through the console or API, and data exports provide external backup options for critical cost analysis workflows.

Backup Capabilities

Cost Explorer does not have traditional backup capabilities as it is a managed analytics service that sources data from AWS billing systems. The service provides data export functionality through CSV downloads and API access for external data preservation. Custom reports and dashboard configurations can be saved within the service, and organizations can implement their own data backup strategies using the Cost Explorer API to programmatically extract and archive cost data externally.

Disaster Recovery

Cost Explorer is a fully managed service with inherent disaster recovery built into AWS infrastructure. The service does not require separate disaster recovery planning as it automatically maintains high availability and data consistency across AWS regions:

- Cost data is sourced from resilient AWS billing systems, ensuring continuous availability
- Organizations can maintain business continuity through API access and exported data, allowing cost analysis to continue even during localized service interruptions

Recovery Objectives

Cost Explorer supports rapid recovery objectives through its cloud-native architecture and API-first design. The service maintains near real-time data availability with 24-hour refresh cycles, supporting low RTO requirements for cost monitoring and analysis. Organizations can achieve minimal RPO through continuous data export and API-based backup strategies. The service's high availability design and integration with AWS's global

infrastructure ensures consistent access to critical cost data for business continuity and financial decision-making processes.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/cost-management/latest/userguide/management-limits.html

FAQ: https://aws.amazon.com/aws-cost-management/aws-cost-explorer/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/cost-management/latest/userguide/ce-what-is.html

User Guide: https://docs.aws.amazon.com/cost-management/latest/userguide/ce-what-is.html

API Reference: https://docs.aws.amazon.com/aws-cost-management/latest/APIReference/API_Operations_AWS_Cost_Explorer_Service.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/aws-cost-management/aws-cost-explorer/pricing/

Cost Optimization

Cost Explorer offers unique cost optimization through automated Reserved Instance and Savings Plans recommendations based on actual usage patterns. The service provides rightsizing recommendations for EC2 instances and identifies underutilized resources across all AWS services:

- Its AI-powered forecasting helps predict optimal commitment levels for cost savings, while cost anomaly detection prevents unexpected spending
- The platform enables granular cost allocation through tagging strategies and supports automated budget alerts to prevent overspending
- Organizations can leverage the Cost Comparison feature to identify cost drivers and implement data-driven optimization strategies across their entire AWS infrastructure

Savings Considerations

Cost Explorer maximizes savings through comprehensive Reserved Instance and Savings Plans analysis, potentially reducing compute costs by up to 75% through proper

commitment strategies. The service identifies idle and underutilized resources, enabling right-sizing decisions that can reduce costs by 20-50% for oversized instances:

- Its forecasting capabilities help organizations make informed capacity planning decisions, avoiding both over-provisioning and emergency scaling costs
- The automated cost anomaly detection prevents budget overruns and unexpected charges, while detailed usage analytics enable teams to optimize resource allocation and eliminate waste across their AWS environment, resulting in typical cost reductions of 15-30% for well-managed accounts

Service Name

AWS Database Migration Service

Service Overview

AWS Database Migration Service (AWS DMS) is a fully managed cloud service that enables migration of relational databases, data warehouses, NoSQL databases, and other data stores to AWS or between cloud and on-premises setups. AWS DMS simplifies the migration process by automating discovery of source data stores, schema conversion, and data migration. The service supports both one-time migrations and ongoing replication to keep sources and targets synchronized. AWS DMS handles deployment, monitoring, and scaling automatically, offering pay-as-you-go pricing, automatic failover, and secure data encryption both in transit and at rest using AWS KMS and SSL.

Key Use Cases

- Database modernization: Migrating from legacy databases to modern, cost-effective engines like Amazon Aurora or open-source databases
- Cloud migration: Moving on-premises databases to AWS managed services like Amazon RDS with minimal downtime
- Data warehouse migration: Migrating data warehouses to Amazon Redshift for analytics and reporting
- Disaster recovery: Setting up cross-region database replication for business continuity
- Database consolidation: Combining multiple databases into a single, more efficient database infrastructure
- Real-time data replication: Enabling continuous data synchronization between source and target databases for analytics

Features

- **DMS Fleet Advisor**: Automatically inventories on-premises database and analytic servers to build migration inventory
- **DMS Schema Conversion**: Automatically assesses and converts source schemas to target database engine format
- **Change Data Capture (CDC)**: Replicates ongoing changes to keep sources and targets synchronized in real-time
- **Homogeneous Data Migrations**: Serverless migrations between same database engines with automatic resource scaling
- **Heterogeneous Migrations**: Supports migrations between different database engines with schema conversion

- **Automatic Failover**: Provides high availability with backup replication servers taking over seamlessly
- **Data Encryption**: Encrypts data at rest using AWS KMS and in transit using SSL/TLS
- **Multi-AZ Deployment**: Supports replication instances across multiple Availability Zones for high availability

Value Proposition

AWS Database Migration Service provides a fully managed, cost-effective solution that eliminates the complexity of traditional database migration. Customers benefit from automatic infrastructure management, eliminating the need for capacity planning, hardware procurement, and software installation:

- The service offers minimal downtime through continuous data replication, automatic failover capabilities, and enterprise-grade security with encryption
- AWS DMS supports both homogeneous and heterogeneous migrations with built-in schema conversion tools, making it versatile for various migration scenarios
- The pay-as-you-go pricing model ensures cost efficiency, while integration with other AWS services provides a comprehensive migration ecosystem

Service Integration

AWS DMS integrates seamlessly with Amazon RDS and Amazon Aurora as target databases, Amazon S3 for data storage and intermediate migration steps, and Amazon EC2 for hosting replication instances. The service works with AWS CloudFormation for infrastructure as code deployment, AWS Secrets Manager for secure credential storage, and AWS KMS for data encryption:

- Integration with Amazon CloudWatch provides monitoring and alerting capabilities, while Amazon SNS enables notifications
- AWS DMS also connects with AWS Schema Conversion Tool for schema migration, Amazon Redshift for data warehouse migrations, Amazon DynamoDB for NoSQL migrations, and Amazon Kinesis for real-time data streaming
- The service leverages AWS IAM for access control and integrates with AWS Backup for data protection strategies

Onboarding and Offboarding

Customers begin by setting up an AWS account and creating prerequisite resources including VPC, security groups, and target databases. The migration process involves creating a replication instance to host migration tasks, configuring source and target endpoints with connection details, and creating migration tasks specifying data transfer requirements:

- AWS provides comprehensive documentation, step-by-step migration guides, and the AWS DMS console for easy setup
- Customers can use DMS Fleet Advisor to discover source infrastructure and DMS Schema Conversion for schema assessment
- The service includes built-in data validation tools and monitoring capabilities
- AWS offers free tier access with 750 hours of dms.t3.micro instance usage monthly for the first year, enabling customers to test migrations before full implementation

Getting Started Guide:
https://docs.aws.amazon.com/dms/latest/userguide/CHAP_GettingStarted.html

Data Removal

Customers can remove data by deleting replication tasks, which stops ongoing data synchronization, followed by terminating replication instances and deleting associated EBS storage volumes. Target database data must be manually removed from destination systems like Amazon RDS or Amazon S3. AWS DMS provides cleanup procedures for removing endpoints, subnet groups, and certificates. Customers should also delete any DMS Fleet Advisor collectors and migration project resources. CloudFormation templates can automate resource cleanup if used for initial deployment.

Backup/Restore and Disaster Recovery

AWS Database Migration Service itself does not provide direct backup capabilities but enables disaster recovery scenarios through continuous data replication to secondary regions. DMS supports cross-region replication for disaster recovery setups, maintaining synchronized standby databases. The service integrates with AWS Backup for automated database snapshots and AWS Elastic Disaster Recovery for comprehensive DR solutions. Customers can use DMS for real-time data replication to maintain DR sites with minimal RPO and RTO objectives.

Backup Capabilities

AWS DMS does not have built-in backup capabilities and does not directly integrate with AWS Backup. However, target databases created through DMS migrations (such as Amazon RDS instances) can be backed up using AWS Backup or native database backup features. DMS focuses on data migration and replication rather than backup functionality, with backup responsibilities handled by the target database services.

Disaster Recovery

AWS DMS supports disaster recovery through continuous data replication capabilities but does not directly integrate with AWS Elastic Disaster Recovery. DMS enables cross-region database replication for DR scenarios, maintaining synchronized standby databases with minimal recovery point objectives. Organizations can use DMS as part of comprehensive

disaster recovery strategies by replicating data to DR sites and maintaining database synchronization across regions.

Recovery Objectives

AWS DMS helps customers achieve low RPO through continuous change data capture (CDC) that replicates data changes in near real-time, minimizing data loss potential. For RTO objectives, the service enables rapid failover scenarios through maintained synchronized replica databases. Cross-region replication capabilities support disaster recovery with recovery times dependent on failover processes rather than data restoration, enabling organizations to meet aggressive RPO and RTO requirements for business continuity.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/dms/latest/userguide/CHAP_Limits.html

FAQ: https://aws.amazon.com/dms/faqs/

Technical Requirements

Service Documentation:
https://docs.aws.amazon.com/dms/latest/userguide/Welcome.html

User Guide: https://docs.aws.amazon.com/dms/latest/userguide/Welcome.html

API Reference: https://docs.aws.amazon.com/dms/latest/APIReference/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/dms/pricing/

Cost Optimization

AWS DMS offers multiple cost optimization options including serverless capabilities that automatically scale based on workload demand, eliminating over-provisioning costs. Database Savings Plans provide up to 35% savings with one-year commitments:

- The service supports pay-as-you-go pricing for on-demand instances, allowing customers to pay only for migration duration
- Right-sizing replication instances based on actual throughput requirements optimizes costs
- Using parallel loading for large tables and optimized change processing reduces migration time and associated costs

- The AWS Free Tier includes 750 hours monthly of dms.t3.micro usage, enabling cost-effective testing and small migrations

Savings Considerations

Primary cost savings come from eliminating traditional migration infrastructure through fully managed service delivery, reducing hardware procurement and maintenance costs. Automated capacity management and scaling eliminate over-provisioning expenses:

- Minimal downtime capabilities reduce business disruption costs compared to traditional migration approaches
- Database engine modernization enabled by DMS can result in long-term licensing cost reductions by migrating to open-source or AWS-native database services
- Efficient migration tools and automation reduce project timelines and associated labor costs
- Integration with other AWS services creates economies of scale and simplified billing management

Service Name

AWS DataSync

Service Overview

AWS DataSync is an online data transfer service that simplifies and accelerates data migrations to AWS and facilitates data movement between on-premises storage, edge locations, other cloud providers, and AWS Storage services. DataSync automates data-transfer processes and infrastructure requirements for high performance and secure data transfers. It uses a purpose-built network protocol and parallel, multi-threaded architecture to accelerate transfers while ensuring end-to-end security through encryption and data integrity validation. DataSync supports various storage systems including NFS, SMB, HDFS, self-managed object storage, cloud providers like Azure and Google Cloud, and AWS services like S3, EFS, and FSx.

Key Use Cases

- Data Migration: Transferring active datasets from on-premises or other cloud providers to AWS storage services for cloud adoption
- Data Archiving: Moving cold data from on-premises storage to durable and secure long-term storage classes like S3 Glacier Flexible Retrieval or S3 Glacier Deep Archive
- Data Replication: Copying data to AWS for business continuity, disaster recovery, and creating backup copies across regions
- Hybrid Cloud Workflows: Enabling recurring data transfers between on-premises and AWS for ongoing processing and analysis
- Data Lake Building: Transferring large datasets to AWS for analytics, machine learning, and data processing workloads

Features

- **Purpose-built Network Protocol**: Uses AWS-designed transfer protocol for optimized data movement with parallel, multi-threaded architecture
- **Network Optimization**: Performs incremental transfers, in-line compression, sparse file detection, and bandwidth controls
- **Multi-cloud Support**: Supports data movement between on-premises, AWS, and other cloud providers including Azure, Google Cloud, and object storage systems
- **Enhanced and Basic Modes**: Enhanced mode offers higher performance and parallel processing, while Basic mode supports broader compatibility
- **Security and Encryption**: Encrypts data in transit using TLS 1.3, supports Kerberos authentication, and ensures data integrity validation

- **Metadata Preservation**: Preserves file permissions, timestamps, ownership, and other metadata during transfers
- **Scheduling and Automation**: Built-in scheduling mechanism for periodic data transfer tasks and automated transfer management
- **Monitoring and Auditing**: Provides JSON-formatted task reports, CloudWatch integration, and CloudTrail auditing capabilities

Value Proposition

AWS DataSync eliminates the complexity and cost of building custom data transfer solutions or licensing expensive commercial network acceleration software. It provides up to 10 times faster transfer speeds than open-source tools while automating data transfer processes, infrastructure management, and ensuring security:

- DataSync offers a pay-as-you-go pricing model with no upfront costs, supports virtually unlimited numbers of files, and handles encryption, verification, and monitoring automatically
- The service integrates seamlessly with AWS storage services and provides granular bandwidth controls to minimize impact on production systems

Service Integration

DataSync integrates deeply with core AWS storage services including Amazon S3 (all storage classes), Amazon EFS, Amazon FSx for Windows File Server, Amazon FSx for Lustre, Amazon FSx for OpenZFS, and Amazon FSx for NetApp ONTAP. It leverages AWS IAM for access control, AWS Secrets Manager for credential storage, Amazon CloudWatch for monitoring and metrics, AWS CloudTrail for auditing, and supports AWS PrivateLink for private connectivity. DataSync also integrates with AWS Backup for centralized backup policies and works with data analytics services like AWS Glue, Amazon Athena, and Amazon QuickSight for transferred data analysis.

Onboarding and Offboarding

Getting started with DataSync involves signing up for an AWS account, creating an IAM user with appropriate permissions, and choosing the transfer method based on source and destination. For on-premises transfers, customers deploy and activate a DataSync agent on VMware ESXi, KVM, Nutanix AHV, Microsoft Hyper-V, or Amazon EC2:

- For cloud-to-cloud transfers, no agent may be required
- Customers then create source and destination locations, configure transfer tasks with options like bandwidth controls and scheduling, and start the transfer using the DataSync console, CLI, or API

Getting Started Guide: https://docs.aws.amazon.com/datasync/latest/userguide/setting-up.html

Data Removal

DataSync does not store customer data permanently - it only transfers data between source and destination locations. To offboard, customers can delete DataSync tasks, locations, and agents through the console or API. Any transferred data remains in the destination storage service and must be deleted separately according to that service's procedures. Customers should also remove any associated IAM roles and policies created specifically for DataSync operations.

Backup/Restore and Disaster Recovery

DataSync enables backup and disaster recovery by replicating data to AWS storage services across regions. It supports automated replication schedules for data protection and can transfer data to various S3 storage classes including Glacier for long-term archival. DataSync integrates with AWS Backup for centralized backup management and can replicate Amazon EFS and FSx file systems to different AWS regions for disaster recovery purposes.

Backup Capabilities

DataSync provides backup capabilities through scheduled recurring transfers and integrates with AWS Backup service for centralized policy management. It supports transferring data to Amazon S3 storage classes optimized for backup scenarios, including S3 Glacier Instant Retrieval, S3 Glacier Flexible Retrieval, and S3 Glacier Deep Archive. DataSync can perform incremental backups by transferring only changed data.

Disaster Recovery

DataSync supports disaster recovery by enabling cross-region replication of data to AWS storage services. It can replicate Amazon EFS file systems, FSx file systems, and other data to different AWS regions. DataSync does not directly integrate with AWS Elastic Disaster Recovery but serves as a data replication tool that supports broader disaster recovery strategies by ensuring data availability across regions.

Recovery Objectives

DataSync helps customers meet RPO objectives through scheduled incremental transfers that can run as frequently as every hour, minimizing data loss. For RTO objectives, DataSync's high-speed transfer capabilities and automation reduce the time needed to restore data to recovery locations. The service's ability to transfer data at speeds up to 10 times faster than traditional tools helps achieve aggressive recovery time targets.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/datasync/latest/userguide/datasync-limits.html

FAQ: https://aws.amazon.com/datasync/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/datasync/latest/userguide/what-is-datasync.html

User Guide: https://docs.aws.amazon.com/datasync/latest/userguide/setting-up.html

API Reference: https://docs.aws.amazon.com/datasync/latest/userguide/API_Reference.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/datasync/pricing/

Cost Optimization

DataSync offers several cost optimization features including pay-as-you-go pricing with no upfront costs, bandwidth throttling to control network usage, incremental transfers that only move changed data, and built-in compression to reduce data transfer volumes. Enhanced mode provides better performance per dollar with parallel processing. Customers can optimize costs by scheduling transfers during off-peak hours, using appropriate S3 storage classes for destinations, and leveraging AWS Direct Connect to reduce data transfer charges for large migrations.

Savings Considerations

Main cost savings include eliminating the need for expensive commercial network acceleration software, reducing operational overhead through automation, and minimizing network bandwidth usage through compression and incremental transfers. DataSync's speed improvements (up to 10x faster than open-source tools) reduce the time and resources needed for data migration projects. The service's ability to transfer data directly to cost-effective S3 storage classes like Glacier can significantly reduce long-term storage costs compared to maintaining on-premises infrastructure.

Service Name

AWS Data Exchange

Service Overview

AWS Data Exchange is a comprehensive data marketplace service that enables organizations to easily discover, subscribe to, and use third-party data in the AWS Cloud. The service facilitates secure data sharing between data providers and subscribers, offering over 3,500 data sets from more than 300 providers. Data can be accessed through files, APIs, or Amazon Redshift queries and delivered directly to data lakes, applications, analytics, and machine learning models. AWS Data Exchange eliminates the need for providers to build and maintain data delivery, entitlement, or billing infrastructure while providing subscribers with a centralized platform to find, evaluate, and consume external data sets across various industries.

Key Use Cases

- Financial Services: Access market data, ESG ratings, consumer spending patterns, and portfolio analytics for risk management, capital markets analysis, and compliance reporting
- Healthcare and Life Sciences: Accelerate research with medical imaging data, clinical trial information, claims data, and genomic test results to improve patient care and quantify health outcomes
- Marketing and Advertising: Enhance campaigns with consumer behavior data, demographic information, location intelligence, and social media insights to improve targeting and ROI

Features

- **Multi-format Data Support**: Supports Files, API, Amazon Redshift datashares, Amazon S3 data access, and AWS Lake Formation data permissions
- **Automated Data Delivery**: Automatic export of new or updated data revisions to Amazon S3 buckets with structured storage
- **Subscription Management**: Centralized platform to track and manage all data grants and AWS Marketplace subscriptions
- **Data Discovery**: Browse catalog with over 3,500 data sets across multiple industries with filtering and search capabilities
- **Security and Compliance**: End-to-end encryption, subscription verification, IAM-based access control, and compliance with data privacy laws
- **Flexible Pricing Models**: Support for subscription-based, pay-as-you-go, private offers, and custom terms with consolidated billing

Value Proposition

AWS Data Exchange provides unparalleled advantages by eliminating the complexity of third-party data procurement and management. Customers benefit from simplified contracts, consolidated secure billing, and unified data ingestion across multiple providers:

- The service offers native integration with AWS analytics and machine learning services, reducing time-to-insight and operational overhead
- Data providers eliminate the need to build delivery infrastructure while reaching millions of AWS customers, while subscribers gain access to premium data sets with automated updates and built-in security controls, significantly reducing the total cost of ownership for external data initiatives

Service Integration

AWS Data Exchange integrates deeply with core AWS services including Amazon S3 for data storage and export, Amazon API Gateway for API-based data products, Amazon Redshift for direct database queries without ETL processes, and AWS Lake Formation for data lake permissions. The service leverages AWS Marketplace for product discovery and billing, AWS IAM for access control, Amazon Athena for data querying, Amazon SageMaker for machine learning workflows, AWS Glue DataBrew for data preparation, Amazon QuickSight for visualization, and Amazon EMR for big data processing, creating a comprehensive ecosystem for data consumption and analytics.

Onboarding and Offboarding

Getting started with AWS Data Exchange requires an AWS account and appropriate IAM permissions. Subscribers can immediately browse the catalog through the AWS Management Console or AWS Marketplace, evaluate products using data samples and dictionaries, and subscribe with a few clicks:

- Data providers must first register as AWS Marketplace sellers, meet eligibility requirements, and undergo a qualification process
- Both roles can access comprehensive documentation, workshops, and expert guidance
- The service provides step-by-step getting started guides for subscribers and providers, with automated notification systems for subscription updates and new data availability

Getting Started Guide: https://docs.aws.amazon.com/data-exchange/latest/userguide/setting-up.html

Data Removal

When subscriptions end, access to AWS Data Exchange data sets is immediately terminated. However, subscribers retain access to previously exported files in their own Amazon S3 buckets. Data providers can unpublish products from the catalog while existing subscriptions remain active until expiration. Subscribers must review their Data Subscription Agreement to understand requirements for deleting exported data. The service provides clear guidance on data retention obligations and removal procedures.

Backup/Restore and Disaster Recovery

AWS Data Exchange relies on underlying AWS service infrastructure for backup and disaster recovery capabilities. Data stored in Amazon S3 benefits from S3's built-in durability and availability features. The service operates across multiple AWS regions with automatic replication of metadata and subscription information. However, AWS Data Exchange does not provide specific backup services and relies on customers to implement their own data protection strategies for exported data.

Backup Capabilities

AWS Data Exchange does not have native backup capabilities and does not integrate directly with AWS Backup. The service depends on the durability and availability of underlying AWS services like Amazon S3, which provides 99.999999999% (11 9's) durability. Customers are responsible for implementing backup strategies for data exported to their own AWS accounts using standard AWS backup services and best practices.

Disaster Recovery

AWS Data Exchange does not directly support AWS Elastic Disaster Recovery integration. The service achieves resilience through its multi-region architecture and dependency on highly available AWS services:

- Disaster recovery is primarily handled through AWS's infrastructure resilience, with data sets and subscriptions accessible across supported regions
- Customers must implement their own disaster recovery strategies for exported data using appropriate AWS services

Recovery Objectives

AWS Data Exchange helps customers meet RPO and RTO objectives through its distributed architecture across multiple AWS regions and high availability of underlying services. The service provides immediate access to subscribed data sets and supports automated exports to reduce manual recovery processes. However, specific RPO/RTO targets depend on customer implementation of backup and recovery strategies for exported data rather than inherent service capabilities.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/data-exchange/latest/userguide/limits.html

FAQ: https://aws.amazon.com/data-exchange/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/data-exchange/latest/userguide/what-is.html

User Guide: https://docs.aws.amazon.com/data-exchange/latest/userguide/

API Reference: https://docs.aws.amazon.com/data-exchange/latest/apireference/welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/data-exchange/pricing/

Cost Optimization

AWS Data Exchange offers unique cost optimization through its tiered fulfillment fee structure for data providers and consolidated billing for subscribers managing multiple data subscriptions. Providers can optimize costs by leveraging automatic data export features, reducing manual delivery overhead:

- The service supports various pricing models including subscription-based and pay-as-you-go options, allowing customers to align costs with usage patterns
- Integration with AWS native services enables customers to minimize data transfer costs by keeping data within the AWS ecosystem and leveraging services like S3 storage classes and data lifecycle policies

Savings Considerations

Major cost savings come from eliminating the need to build custom data delivery infrastructure, reducing operational overhead through automated data management, and leveraging consolidated billing across multiple data providers. Customers save on integration costs through native AWS service compatibility and reduce time-to-market for data-driven applications:

- The pay-as-you-go model for certain products eliminates upfront commitments, while subscription models provide predictable costs

- Additional savings emerge from reduced data transfer fees by keeping operations within AWS, optimized storage costs through S3 integration, and elimination of manual data procurement and management processes

Service Name

AWS Device Farm

Service Overview

AWS Device Farm is an app testing service that enables developers to test and interact with Android, iOS, and web apps on real, physical phones and tablets hosted by AWS. The service provides two main testing approaches: remote access for real-time interaction through web browsers or automated testing using Appium from local clients, and automated app testing using Device Farm's managed test execution environment. Device Farm supports both native and hybrid mobile applications, providing comprehensive test reports with high-level results, logs, screenshots, and test artifacts to help developers identify and resolve issues across multiple device configurations.

Key Use Cases

- Mobile app testing: Testing Android and iOS native and hybrid apps across multiple real devices simultaneously to ensure compatibility and performance
- Cross-device compatibility validation: Verifying app functionality, visual rendering, and user experience across different screen sizes, operating systems, and device manufacturers
- Automated testing integration: Incorporating Device Farm into CI/CD pipelines for continuous testing and quality assurance during development cycles
- Bug reproduction and debugging: Using remote access sessions to reproduce specific device issues, debug Appium tests, and troubleshoot app behavior in real-time
- Desktop browser testing: Testing web applications on real desktop browsers using Selenium to ensure cross-browser compatibility

Features

- **Real Device Testing**: Access to a wide range of physical Android and iOS devices for authentic testing conditions
- **Remote Access Sessions**: Real-time interaction with devices through web browsers with 150-minute session limits
- **Automated Test Execution**: Parallel test execution across multiple devices using various testing frameworks
- **Managed Appium Endpoint**: Fully-managed Appium endpoint for connecting and running interactive tests from local IDEs with live video and log streaming
- **Multiple Test Framework Support**: Support for Appium, XCTest, XCTest UI, Instrumentation, and built-in fuzz tests

- **Private Device Fleet**: Dedicated physical devices deployed exclusively for specific AWS accounts in secure data centers
- **Desktop Browser Testing**: Testing web applications on real desktop browsers using Selenium
- **Test Reporting and Analytics**: Comprehensive reports with pass/fail information, crash reports, logs, screenshots, and performance data retained for 400 days
- **VPC Integration**: Secure connectivity to private endpoints and on-premises applications through VPC-ENI
- **Environment Variables and IAM Roles**: Support for environment variables for test configuration and IAM role integration for accessing other AWS services

Value Proposition

AWS Device Farm eliminates the need to purchase, maintain, and manage physical device labs, significantly reducing infrastructure costs and operational overhead. The service provides access to a constantly updated fleet of real devices reflecting current market conditions, ensuring comprehensive testing coverage without device procurement delays:

- Device Farm's integration with AWS services enables seamless incorporation into existing development workflows and CI/CD pipelines
- The managed Appium endpoint and real-time debugging capabilities accelerate development cycles by providing immediate feedback
- Pay-as-you-go pricing ensures customers only pay for actual usage, while private device options offer dedicated resources for sensitive applications requiring enhanced security and compliance

Service Integration

Device Farm integrates deeply with AWS CodePipeline for automated testing in CI/CD workflows, enabling continuous integration and deployment processes. The service connects with Amazon S3 for storing test artifacts, logs, and application files, while Amazon CloudWatch provides monitoring and logging capabilities for test execution:

- IAM roles enable secure access to other AWS resources during test execution
- VPC integration allows secure connectivity to private endpoints and on-premises applications
- AWS Lambda and Step Functions can be used for custom test workflows and automation
- The service also integrates with AWS Backup for data protection strategies and supports connection to Amazon DynamoDB and other AWS services through IAM role-based access during test execution

Onboarding and Offboarding

Getting started with AWS Device Farm requires creating an AWS account and setting up an IAM user with appropriate Device Farm permissions. Customers begin by creating a project in the Device Farm console, then can either upload their mobile application (.apk or .ipa files) or select from sample applications:

- Next, they choose their preferred testing approach: automated testing with supported frameworks (Appium, XCTest, Instrumentation) or remote access sessions for manual testing
- Device selection involves creating device pools with specific criteria or using curated device sets
- The service provides a free tier for initial evaluation, after which customers can purchase device slots based on their testing requirements and scale as needed

Getting Started Guide:
https://docs.aws.amazon.com/devicefarm/latest/developerguide/getting-started.html

Data Removal

AWS Device Farm automatically manages data lifecycle with built-in retention policies. Apps and test files are stored for 30 days before automatic deletion, while session logs and captured videos are retained for 400 days. Customers can manually delete their files at any time through the console or API. For private devices, AWS ensures secure destruction of hardware when terminating the service. Public devices undergo comprehensive cleanup including app uninstallation and factory reset between sessions. Customers maintain control over their data and can implement additional backup strategies if longer retention is required.

Backup/Restore and Disaster Recovery

AWS Device Farm provides limited native backup capabilities, focusing primarily on automatic data retention policies rather than comprehensive backup solutions. The service is only available in the us-west-2 region, making it essential for customers to implement their own backup and recovery processes rather than relying solely on Device Farm for uploaded content. Test artifacts, logs, and videos are stored with automatic lifecycle management, but customers should not consider Device Farm as their sole source of critical test assets and should maintain separate backup strategies for important applications and test suites.

Backup Capabilities

Device Farm does not integrate with AWS Backup service and has limited native backup functionality. The service automatically retains test artifacts, session videos, and logs for defined periods (30 days for files, 400 days for logs and videos) but does not provide comprehensive backup solutions:

- Customers are responsible for backing up their application files, test scripts, and any critical test data outside of Device Farm
- The service's automatic data lifecycle management serves more as data retention rather than true backup capability

Disaster Recovery

AWS Device Farm does not integrate with AWS Elastic Disaster Recovery and has limited disaster recovery capabilities due to its single-region deployment in us-west-2. The service relies on AWS's underlying infrastructure resilience within that region, including multiple Availability Zones for fault tolerance:

- However, customers cannot implement cross-region disaster recovery strategies for Device Farm resources
- Organizations requiring disaster recovery should implement their own backup processes and maintain alternative testing strategies in case of regional service disruptions

Recovery Objectives

Device Farm's contribution to RPO and RTO objectives is limited due to its single-region architecture and focus on testing rather than production workloads. The service can help customers validate application resilience and disaster recovery procedures through comprehensive testing across multiple device configurations. However, for customers' own recovery objectives, Device Farm's automatic data retention (30-400 days) and regional concentration mean organizations must implement separate backup and recovery strategies for critical testing assets and maintain alternative testing approaches to meet aggressive RPO/RTO requirements during service disruptions.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/general/latest/gr/devicefarm.html

FAQ: https://aws.amazon.com/device-farm/faqs/

Technical Requirements

Service Documentation:
https://docs.aws.amazon.com/devicefarm/latest/developerguide/welcome.html

User Guide:
https://docs.aws.amazon.com/devicefarm/latest/developerguide/welcome.html

API Reference:
https://docs.aws.amazon.com/devicefarm/latest/APIReference/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/device-farm/pricing/

Cost Optimization

AWS Device Farm offers several cost optimization strategies specific to mobile and web testing. The pay-as-you-go pricing model allows customers to pay only for actual device usage time, eliminating upfront hardware investments and maintenance costs:

- Device slot purchasing enables bulk pricing for frequent testing needs while avoiding per-minute charges
- Desktop browser testing uses a simple per-instance-minute pricing model at $0.005 per minute
- Customers can optimize costs by efficiently managing test duration, using device pools strategically to target specific device combinations, and leveraging the 150-minute session limit effectively
- Private devices provide dedicated resources without additional per-usage fees once deployed

Savings Considerations

The primary cost savings with AWS Device Farm come from eliminating the need to purchase, maintain, and update physical device labs, which can cost hundreds of thousands of dollars annually. Organizations save on facility costs, device procurement, maintenance staff, and continuous hardware refresh cycles:

- The service's managed infrastructure eliminates IT overhead for device management, software updates, and testing environment maintenance
- Pay-as-you-go pricing prevents over-provisioning of testing resources, while device slot options provide predictable costs for regular testing workflows
- Integration with CI/CD pipelines reduces manual testing time and associated labor costs
- The free tier allows evaluation without initial investment, and customers can scale testing capacity precisely to match development needs

Service Name

AWS Directory Service

Service Overview

AWS Directory Service provides multiple ways to use Microsoft Active Directory (AD) and Lightweight Directory Access Protocol (LDAP) with other AWS services. It is a collection of directory options that store information about users, groups, and devices, enabling administrators to manage access to information and resources. The service includes three main options: AWS Managed Microsoft AD (a fully managed Microsoft Windows Server Active Directory service in the AWS Cloud), AD Connector (a proxy service for connecting existing on-premises Microsoft AD to AWS), and Simple AD (a Samba 4-based directory compatible with basic Active Directory features). The service enables migration of Active Directory-aware applications to the AWS Cloud while supporting various AWS managed applications and services.

Key Use Cases

- Enterprise directory services: Migrate Active Directory-aware applications like Microsoft SharePoint, SQL Server, and .NET applications to AWS Cloud
- Hybrid cloud integration: Extend existing on-premises Active Directory infrastructure to AWS using AD Connector or trust relationships
- Single Sign-On (SSO): Enable users to access AWS Management Console and applications using existing Active Directory credentials
- AWS service authentication: Authenticate and authorize access for AWS services like WorkSpaces, WorkDocs, RDS for SQL Server, Amazon Chime, and Amazon Connect
- Compliance workloads: Support applications subject to HIPAA and PCI DSS compliance requirements
- Multi-region deployments: Replicate directories across multiple AWS regions for global workload management

Features

- **AWS Managed Microsoft AD**: Fully managed Windows Server 2019 Active Directory with high availability, automatic patching, monitoring, and recovery across multiple Availability Zones
- **AD Connector**: Proxy service that connects AWS applications to existing on-premises Microsoft Active Directory without caching information in the cloud
- **Simple AD**: Samba 4-based directory providing basic Active Directory compatibility for low-scale, cost-effective directory needs

- **Multi-Region Replication**: Enterprise Edition feature enabling directory replication across multiple AWS regions for global workload support
- **Directory Sharing**: Share directories across multiple AWS accounts and VPCs within an organization
- **Trust Relationships**: Create trust relationships between AWS Managed Microsoft AD and existing on-premises or cloud-based Active Directory domains
- **Schema Extensions**: Extend Active Directory schema to support custom attributes and object classes
- **LDAPS Support**: Secure LDAP communications with server-side and client-side LDAPS encryption
- **Multi-Factor Authentication**: Integration with RADIUS servers for MFA support across directory types

Value Proposition

AWS Directory Service offers significant advantages over self-managed Active Directory solutions by providing a fully managed service that eliminates the operational overhead of maintaining domain controllers, patching, monitoring, and ensuring high availability. The service delivers enhanced security with built-in encryption, compliance certifications (HIPAA, PCI DSS), and AWS security best practices:

- It provides seamless integration with over 15 AWS services and applications, enabling rapid cloud adoption without requiring directory infrastructure changes
- The service offers flexible deployment options including standalone cloud directories, hybrid connectivity to on-premises infrastructure, and multi-region replication for global scalability
- Cost optimization is achieved through pay-as-you-use pricing, elimination of hardware costs, and reduced administrative overhead while maintaining enterprise-grade reliability and performance

Service Integration

AWS Directory Service integrates deeply with core AWS services including Amazon WorkSpaces for virtual desktops, Amazon WorkDocs for document collaboration, Amazon WorkMail for email services, and Amazon Chime for communications. It supports database services through Amazon RDS for SQL Server, Oracle, and PostgreSQL with Windows Authentication:

- The service enables single sign-on to the AWS Management Console through AWS IAM Identity Center integration and supports Amazon EC2 domain joining for both Windows and Linux instances
- Additional integrations include Amazon FSx for Windows File Server for shared storage, AWS Client VPN for secure remote access, Amazon Connect for contact

center solutions, AWS Transfer Family for secure file transfers, and AWS Private Certificate Authority for certificate management
- The service also works with AWS Organizations for cross-account directory sharing and AWS CloudTrail for comprehensive logging and monitoring

Onboarding and Offboarding

Customers begin by selecting the appropriate directory type based on their needs: AWS Managed Microsoft AD for full Active Directory features, AD Connector for hybrid connectivity, or Simple AD for basic requirements. The setup process requires creating an Amazon VPC with at least two subnets in different Availability Zones:

- For AWS Managed Microsoft AD, customers provide directory information including DNS name, NetBIOS name, administrator password, and VPC configuration
- The service automatically creates domain controllers and DNS services within 20-40 minutes
- Customers can then join EC2 instances to the domain, configure trust relationships with existing infrastructure, enable AWS applications, and set up user management through the AWS Management Console or Active Directory tools
- A 30-day free trial with 1,500 hours of usage is available for evaluation

Getting Started Guide: https://docs.aws.amazon.com/directoryservice/latest/admin-guide/ms_ad_getting_started.html

Data Removal

When offboarding from AWS Directory Service, customers must first deauthorize all AWS applications and services connected to the directory, remove any trust relationships with external domains, and unjoin all EC2 instances from the domain. The directory deletion process permanently removes all directory data including users, groups, organizational units, and configuration settings. For AWS Managed Microsoft AD, customers should create manual snapshots before deletion for backup purposes, though snapshots have a maximum age limit of 180 days. Once a directory is deleted, all associated data is permanently removed and cannot be recovered.

Backup/Restore and Disaster Recovery

AWS Directory Service provides automated daily snapshots for AWS Managed Microsoft AD directories with a retention period determined by AWS. Customers can create up to 5 manual snapshots per directory for point-in-time recovery, with snapshots having a maximum supported age of 180 days. Directory restoration from snapshots moves the directory back in time and may result in data loss for changes made after the snapshot date. The service includes built-in high availability across multiple Availability Zones and automatic recovery capabilities. However, snapshots cannot be created for AD Connector directories since they only proxy requests to existing infrastructure.

Backup Capabilities

AWS Directory Service supports manual snapshot creation for AWS Managed Microsoft AD directories, allowing up to 5 snapshots per directory with 180-day maximum age. The service includes automated daily snapshots managed by AWS:

- While AWS Directory Service does not directly integrate with AWS Backup service, customers can implement backup strategies for connected resources and applications
- AD Connector and Simple AD have limited backup capabilities since AD Connector proxies to existing infrastructure and Simple AD data recovery relies on directory recreation

Disaster Recovery

AWS Directory Service provides built-in disaster recovery through multi-Availability Zone deployment of domain controllers, ensuring service continuity if one zone fails. Enterprise Edition supports multi-region replication for geographic disaster recovery:

- The service does not directly integrate with AWS Elastic Disaster Recovery, but supports disaster recovery strategies through snapshot restoration, cross-region replication, and integration with other AWS disaster recovery services
- Trust relationships and hybrid connectivity options provide additional resilience by maintaining connections to on-premises infrastructure as backup authentication sources

Recovery Objectives

AWS Directory Service helps customers achieve Recovery Point Objectives (RPO) through manual snapshots that can be created as frequently as needed within the 5-snapshot limit, providing point-in-time recovery capabilities. Recovery Time Objectives (RTO) are supported through built-in high availability across multiple Availability Zones for automatic failover, and multi-region replication in Enterprise Edition for faster geographic recovery. The service's automated recovery mechanisms and managed infrastructure reduce RTO compared to self-managed solutions, while snapshot restoration capabilities enable RPO management based on backup frequency requirements.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/general/latest/gr/ds_region.html

FAQ: https://aws.amazon.com/directoryservice/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/directoryservice/latest/admin-guide/what_is.html

User Guide: https://docs.aws.amazon.com/directoryservice/latest/admin-guide/what_is.html

API Reference: https://docs.aws.amazon.com/directoryservice/latest/devguide/welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/directoryservice/pricing/

Cost Optimization

AWS Directory Service offers several unique cost optimization opportunities including right-sizing directory editions based on object count and feature requirements (Standard vs Enterprise vs Hybrid), optimizing domain controller count by monitoring CloudWatch performance metrics, and leveraging directory sharing to reduce costs across multiple AWS accounts rather than creating separate directories. The 30-day free trial provides 1,500 hours for evaluation without upfront costs:

- Customers can reduce data transfer costs by avoiding NETLOGON and SYSVOL shares for file storage in multi-region deployments
- Simple AD and AD Connector offer no-cost options when used with WorkSpaces, WorkDocs, or WorkMail with minimum user requirements, providing significant savings for basic directory needs

Savings Considerations

Primary cost savings include eliminating on-premises directory infrastructure costs (hardware, software licenses, maintenance), reducing administrative overhead through fully managed services, and paying only for actual usage without upfront commitments. The service eliminates Windows Server licensing costs for domain controllers and reduces operational expenses through automated patching, monitoring, and high availability management:

- Multi-account directory sharing prevents duplicate directory costs across organizations
- Customers can achieve additional savings by choosing appropriate directory types (Simple AD for basic needs, AD Connector for hybrid scenarios), optimizing domain controller counts based on actual performance requirements, and leveraging the free tier benefits when using directories with qualifying AWS applications like WorkSpaces with minimum user thresholds

Service Name

AWS Direct Connect

Service Overview

AWS Direct Connect is a network service that establishes dedicated network connections from on-premises environments to AWS, bypassing the public internet. It enables organizations to create private, direct connections to AWS services using standard Ethernet fiber-optic cables connected to AWS Direct Connect locations. The service provides consistent network performance, reduced latency, increased bandwidth, and enhanced security for hybrid cloud architectures. Direct Connect supports multiple connection types including dedicated connections (1-400 Gbps) and hosted connections (50 Mbps-10 Gbps) through AWS partners, with virtual interfaces enabling access to public AWS services, Amazon VPC, and AWS Transit Gateway across multiple regions.

Key Use Cases

- Enterprise hybrid cloud connectivity: Connecting on-premises data centers to AWS for seamless hybrid architectures with consistent network performance
- High-bandwidth data transfer: Large-scale data migration, backup, and synchronization requiring reliable throughput and reduced data transfer costs
- Low-latency applications: Real-time applications, financial trading systems, and media streaming requiring predictable network performance
- Regulated industries: Organizations requiring private connections due to compliance, security policies, or restrictions on public internet usage
- Multi-region architecture: Global enterprises connecting multiple locations through AWS backbone using SiteLink for site-to-site connectivity
- Disaster recovery: Establishing reliable network paths for backup and disaster recovery solutions with AWS Elastic Disaster Recovery

Features

- **Multiple Connection Types**: Dedicated connections (1-400 Gbps) and hosted connections (50 Mbps-10 Gbps) through AWS partners
- **Virtual Interfaces (VIFs)**: Private, public, and transit VIFs for accessing VPCs, public AWS services, and Transit Gateways
- **Direct Connect Gateway**: Globally available resource enabling connections to multiple VPCs and Transit Gateways across regions
- **SiteLink**: Private site-to-site connectivity between Direct Connect locations using AWS global backbone
- **Encryption Options**: MACsec encryption for dedicated connections and IPsec VPN support for additional security

- **Link Aggregation Groups (LAG)**: Aggregating multiple connections for increased bandwidth and redundancy
- **BGP Routing**: Dynamic routing with BGP and BGP MD5 authentication for reliable path selection
- **Global Availability**: Available at numerous locations worldwide with speeds up to 400 Gbps

Value Proposition

AWS Direct Connect offers significant advantages over internet-based connections including reduced network costs through dedicated bandwidth and lower data transfer charges, enhanced security via private connections that bypass the public internet, improved performance with consistent low latency and higher throughput, and greater reliability with predictable network behavior. Organizations benefit from seamless integration with AWS services, global reach through AWS backbone infrastructure, and flexible deployment options. Direct Connect is particularly valuable for enterprises requiring hybrid cloud architectures, regulated industries with strict security requirements, applications needing consistent performance, and organizations with high-bandwidth data transfer needs where internet connectivity proves inadequate or cost-prohibitive.

Service Integration

AWS Direct Connect integrates deeply with core AWS networking and compute services. It connects directly to Amazon VPC through private virtual interfaces, enabling secure access to EC2 instances, RDS databases, and other VPC resources:

- Integration with AWS Transit Gateway enables scalable multi-VPC connectivity across regions and accounts
- Public virtual interfaces provide access to all AWS public services including S3, CloudFront, and API Gateway
- Direct Connect works with AWS Site-to-Site VPN for encrypted connections, AWS Cloud WAN for unified network management, and AWS Outposts for hybrid edge infrastructure
- It supports AWS Backup for data protection, integrates with Route 53 for DNS resolution, and enhances AWS Elastic Disaster Recovery deployments with reliable private connectivity for mission-critical workloads

Onboarding and Offboarding

Customers begin by assessing network requirements and selecting an appropriate AWS Direct Connect location near their premises or through an AWS Direct Connect Partner. The process involves ordering a connection through the AWS Management Console using the Connection wizard, selecting bandwidth requirements (50 Mbps to 400 Gbps), and choosing resiliency levels (Development/Test, High Resiliency, or Maximum Resiliency):

- After connection approval, customers download the Letter of Authorization and Connecting Facility Assignment (LOA-CFA), coordinate with their network provider or colocation facility for cross-connect installation, configure virtual interfaces for accessing AWS services, and establish BGP sessions
- AWS provides comprehensive documentation, tutorials, and support throughout the setup process to ensure successful deployment

Getting Started Guide:
https://docs.aws.amazon.com/directconnect/latest/UserGuide/Welcome.html

Data Removal

Direct Connect offboarding involves deleting virtual interfaces, terminating BGP sessions, and removing physical cross-connects at Direct Connect locations. Customers must coordinate with network providers to disconnect physical cables and return any leased equipment. Data removal is handled through standard AWS service deletion processes for connected resources like VPCs and S3 buckets. The physical Direct Connect connection itself contains no customer data, as it only provides network connectivity. Organizations should follow their internal data governance procedures for any connected AWS services before terminating Direct Connect connections.

Backup/Restore and Disaster Recovery

Direct Connect enhances disaster recovery capabilities by providing reliable, private network connectivity for backup and recovery operations. It integrates with AWS Elastic Disaster Recovery to establish secure replication channels between on-premises and AWS environments. The service supports high-bandwidth data transfers for backup operations to services like S3 and AWS Backup. Direct Connect enables consistent network performance for disaster recovery scenarios, reducing recovery time objectives (RTO) through predictable connectivity. However, Direct Connect itself does not provide backup capabilities - it serves as the reliable network foundation enabling other AWS backup and disaster recovery services.

Backup Capabilities

Direct Connect does not provide native backup capabilities as it is a network connectivity service. It does not integrate directly with AWS Backup service since there is no data to back up - Direct Connect only provides network transport:

- However, it significantly enhances backup operations by providing dedicated, high-bandwidth connections for data transfer to AWS storage services
- Organizations use Direct Connect to reliably transfer backup data to S3, EBS snapshots, and other AWS storage services with consistent performance and reduced costs compared to internet-based transfers

Disaster Recovery

Direct Connect supports disaster recovery strategies but does not directly integrate with AWS Elastic Disaster Recovery as a protected resource. Instead, it serves as critical network infrastructure enabling reliable disaster recovery implementations:

- Direct Connect provides the stable, high-bandwidth connections necessary for continuous data replication, failover operations, and recovery procedures
- Organizations implement disaster recovery over Direct Connect by establishing redundant connections across multiple locations, using multiple virtual interfaces for different recovery scenarios, and combining Direct Connect with VPN connections for encrypted backup paths when required

Recovery Objectives

Direct Connect helps customers meet Recovery Point Objective (RPO) and Recovery Time Objective (RTO) requirements by providing consistent, high-bandwidth network connectivity that reduces data replication lag and accelerates recovery operations. The reliable network performance enables more frequent backup windows and faster data synchronization, improving RPO. During disaster events, Direct Connect's predictable connectivity reduces network-related delays in failover procedures, helping achieve better RTO. Multiple connection options and redundancy features minimize single points of failure, ensuring disaster recovery systems remain accessible when needed most.

Service Constraints

Service Quotas:
https://docs.aws.amazon.com/directconnect/latest/UserGuide/limits.html

FAQ: https://aws.amazon.com/directconnect/faqs/

Technical Requirements

Service Documentation:
https://docs.aws.amazon.com/directconnect/latest/UserGuide/index.html

User Guide:
https://docs.aws.amazon.com/directconnect/latest/UserGuide/Welcome.html

API Reference:
https://docs.aws.amazon.com/directconnect/latest/APIReference/index.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/directconnect/pricing/

Cost Optimization

Direct Connect offers several cost optimization strategies including right-sizing connections based on actual bandwidth utilization patterns, using hosted connections for lower bandwidth requirements to reduce port hour costs, implementing data transfer optimization by routing traffic through Direct Connect instead of internet gateways to reduce data transfer charges, and leveraging SiteLink for site-to-site connectivity to eliminate multiple internet circuits. Organizations can optimize costs by consolidating multiple connections through Link Aggregation Groups (LAG), using Direct Connect Gateway to share connections across multiple VPCs and accounts, and implementing traffic engineering to maximize utilization of existing connections before adding capacity.

Savings Considerations

Primary cost savings come from reduced data transfer charges compared to internet-based transfers, elimination of multiple internet circuits through consolidated Direct Connect connectivity, and predictable monthly costs through port hour pricing models. Organizations achieve significant savings on high-volume data transfers, reduced telecommunications costs by replacing multiple internet connections with consolidated Direct Connect access, and operational cost reductions through simplified network management:

- Additional savings include reduced need for VPN licensing and management overhead, lower latency enabling more efficient application performance, and the ability to negotiate better rates with telecommunications providers for dedicated circuits
- Enterprise customers often see 20-50% reduction in overall network costs when properly implementing Direct Connect for hybrid architectures

Service Name

AWS Elastic Beanstalk

Service Overview

AWS Elastic Beanstalk is a platform-as-a-service offering that simplifies deploying, managing, and scaling web applications and services. It abstracts infrastructure complexity by automatically handling provisioning of Amazon EC2 instances, load balancing, auto-scaling, and application health monitoring. Supporting multiple programming languages including Java, .NET, PHP, Node.js, Python, Ruby, Go, and Docker containers, Elastic Beanstalk enables developers to focus on writing code while AWS manages the underlying infrastructure. The service automatically provisions necessary AWS resources including servers, databases, load balancers, and auto scaling groups while maintaining full control and visibility over the AWS resources powering the application.

Key Use Cases

- Web application deployment: Rapid deployment of web applications across multiple programming languages and frameworks without infrastructure management
- DevOps acceleration: Streamlined CI/CD workflows with blue/green deployments, application versioning, and rollback capabilities
- Microservices architecture: Deploying containerized applications and APIs with automatic scaling and load balancing
- Development environment provisioning: Quick setup of development, testing, and staging environments with consistent configurations
- Legacy application modernization: Migrating existing applications to cloud with minimal code changes while gaining scalability benefits

Features

- **Multi-platform Support**: Supports Java, .NET, PHP, Node.js, Python, Ruby, Go, and Docker containers with pre-configured runtime environments
- **Automatic Scaling**: Built-in auto-scaling capabilities using Amazon EC2 Auto Scaling with CloudWatch metrics and triggers
- **Load Balancing**: Integrated Elastic Load Balancing with support for Application, Network, and Classic Load Balancers
- **Health Monitoring**: Comprehensive application health monitoring with automatic instance replacement and detailed health dashboards
- **Application Versioning**: Complete application lifecycle management with version control, rollback capabilities, and blue/green deployments

- **Configuration Management**: Centralized configuration templates and integration with AWS Systems Manager Parameter Store and Secrets Manager
- **Security Integration**: Native IAM integration, VPC support, and compliance with HIPAA, SOC, and PCI DSS standards

Value Proposition

Elastic Beanstalk delivers significant time-to-market advantages by eliminating infrastructure management complexity while maintaining full control over AWS resources. Unlike traditional hosting solutions, it provides automatic scaling, built-in monitoring, and seamless integration with the broader AWS ecosystem:

- The service offers zero additional cost beyond underlying AWS resources, making it highly cost-effective
- Key differentiators include support for multiple deployment strategies (rolling, blue/green, immutable), comprehensive platform coverage, and the ability to customize underlying infrastructure when needed
- This makes it ideal for teams wanting rapid deployment capabilities without sacrificing flexibility or AWS native integrations

Service Integration

Elastic Beanstalk deeply integrates with core AWS services to provide a comprehensive application platform. Primary integrations include Amazon EC2 for compute instances, Auto Scaling for automatic capacity management, and Elastic Load Balancing for traffic distribution:

- The service leverages Amazon S3 for application version storage, CloudWatch for monitoring and logging, and CloudFormation for infrastructure orchestration
- Additional integrations encompass Amazon RDS for databases, Amazon VPC for networking, IAM for security, AWS Systems Manager for configuration management, and AWS Secrets Manager for credential storage
- This extensive integration enables Elastic Beanstalk to provide enterprise-grade capabilities while maintaining simplicity in deployment and management workflows

Onboarding and Offboarding

Getting started with Elastic Beanstalk involves creating an AWS account and accessing the service through the AWS Console, CLI, or EB CLI. The process begins by creating an application, uploading application code (or using sample applications), and launching an environment:

- Elastic Beanstalk provides pre-built sample applications for immediate deployment and testing
- Users can deploy applications via ZIP files, Git integration, or container images

- The service automatically provisions necessary AWS resources including EC2 instances, security groups, load balancers, and CloudFormation stacks
- Configuration options allow customization of instance types, scaling policies, and environment variables during the setup process

Getting Started Guide:
https://docs.aws.amazon.com/elasticbeanstalk/latest/dg/GettingStarted.html

Data Removal

Offboarding involves terminating Elastic Beanstalk environments and deleting applications through the console, CLI, or API. The termination process automatically removes all associated AWS resources including EC2 instances, load balancers, Auto Scaling groups, and security groups. Application versions stored in S3 can be manually deleted or configured for automatic deletion through lifecycle policies. Complete data removal requires deleting all application versions, terminating all environments, and finally deleting the application container to ensure no residual resources remain.

Backup/Restore and Disaster Recovery

Elastic Beanstalk provides backup capabilities through application version storage in Amazon S3 and configuration templates. Application code and configurations are automatically preserved in S3, enabling rollback and recovery scenarios. The service integrates with AWS Backup for automated backup of underlying EC2 instances and EBS volumes. Disaster recovery is supported through multi-Availability Zone deployments, cross-region application deployment capabilities, and integration with AWS Elastic Disaster Recovery for comprehensive business continuity planning.

Backup Capabilities

Elastic Beanstalk maintains automatic backups of application versions in Amazon S3 with configurable lifecycle policies. The service preserves up to 1000 application versions by default and supports integration with AWS Backup for underlying infrastructure backup:

- Configuration templates provide backup of environment settings and can be used to recreate environments
- Application source bundles are retained in S3 unless explicitly deleted, providing inherent backup capabilities for application code and configurations

Disaster Recovery

Disaster recovery support includes multi-Availability Zone deployments for high availability and cross-region deployment capabilities for geographic redundancy. The service integrates with AWS Elastic Disaster Recovery for comprehensive DR scenarios and supports blue/green deployment strategies for minimizing downtime during updates. Auto

Scaling and health monitoring provide automatic recovery from instance failures, while configuration templates enable rapid environment recreation in alternate regions.

Recovery Objectives

Elastic Beanstalk helps achieve low RTO through automated scaling, health monitoring, and blue/green deployments that minimize downtime during updates and failures. RPO objectives are supported through continuous application version storage in S3 and regular configuration snapshots. Multi-AZ deployments provide sub-minute failover capabilities, while cross-region deployment strategies enable recovery from regional outages. The combination of automated infrastructure management and AWS integration capabilities supports both aggressive and conservative recovery objectives.

Service Constraints

Service Quotas:
https://docs.aws.amazon.com/elasticbeanstalk/latest/api/API_ResourceQuotas.html

FAQ: https://aws.amazon.com/elasticbeanstalk/faqs/

Technical Requirements

Service Documentation:
https://docs.aws.amazon.com/elasticbeanstalk/latest/dg/Welcome.html

User Guide: https://docs.aws.amazon.com/elasticbeanstalk/latest/dg/Welcome.html

API Reference: https://docs.aws.amazon.com/elasticbeanstalk/latest/api/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/elasticbeanstalk/pricing/

Cost Optimization

Elastic Beanstalk offers unique cost optimization through its zero-service-fee model where customers pay only for underlying AWS resources. Cost optimization strategies include leveraging Spot Instances for significant savings on non-critical environments, implementing time-based scaling to match traffic patterns, and using AWS Graviton processors for improved price-performance ratios:

- The service supports Savings Plans and Reserved Instances for predictable workloads

- Environment-specific optimizations include right-sizing instance types based on application requirements, configuring auto-scaling policies to minimize over-provisioning, and utilizing single-instance environments for development scenarios
- Application version lifecycle policies help manage S3 storage costs by automatically cleaning up old versions

Savings Considerations

Primary cost savings come from Elastic Beanstalk's automatic scaling capabilities that eliminate over-provisioning and reduce waste during low-traffic periods. The service's integration with AWS Cost Optimization tools enables identification of rightsizing opportunities and unused resources:

- Significant savings are achievable through strategic use of Spot Instances for development and testing environments, implementing scheduled scaling for predictable traffic patterns, and leveraging multi-AZ deployments only when high availability is required
- Cross-region deployment strategies should balance disaster recovery requirements with data transfer costs
- Regular review of application versions and cleanup of unused resources prevents unnecessary S3 storage charges

Service Name

AWS Elastic Disaster Recovery

Service Overview

AWS Elastic Disaster Recovery (DRS) is a service that minimizes downtime and data loss by enabling fast, reliable recovery of on-premises and cloud-based applications using block-level replication. It continuously replicates applications and data to a low-cost staging area in AWS, allowing organizations to recover applications running on physical servers, virtual machines, and cloud instances. During a disaster, DRS converts replicated servers to run natively on AWS within minutes, using the most up-to-date server state or previous point-in-time snapshots for recovery.

Key Use Cases

- Disaster recovery: Protect on-premises, AWS EC2, and third-party cloud workloads with continuous replication and rapid failover to AWS
- Ransomware recovery: Launch clean, unencrypted versions of servers from point-in-time snapshots taken before ransomware attacks
- Cross-region disaster recovery: Replicate applications between AWS regions for protection against regional outages
- Cross-availability zone recovery: Enhance business continuity by replicating workloads within different availability zones
- Migration and modernization: Use DRS for one-time migration of workloads from on-premises to AWS cloud infrastructure

Features

- **Block-level Continuous Replication**: Provides near real-time data replication using block-level replication technology with RPOs of seconds
- **Point-in-Time Recovery**: Creates automated snapshots every 10 minutes for the last hour, hourly for 24 hours, and daily for 7 days
- **Launch Management Settings**: Allows configuration of default launch settings, bulk modifications, and recovery availability zone selection
- **Post-Launch Actions**: Enables automated actions after recovery instance launch using AWS Systems Manager commands
- **Automated Network Replication**: Replicates AWS network components including subnets, security groups, and route tables
- **Drill Testing**: Provides non-disruptive testing capabilities to validate disaster recovery plans without impacting production

Value Proposition

AWS Elastic Disaster Recovery offers superior value through its staging area design that reduces costs by eliminating idle recovery site resources while maintaining continuous protection. It provides faster recovery times with RTOs measured in minutes and RPOs in seconds, compared to traditional disaster recovery solutions:

- The service automatically converts servers to run natively on AWS during recovery, supports multiple source environments (on-premises, AWS, third-party clouds), and includes comprehensive testing capabilities
- Its pay-per-source-server pricing model at $0.028 per hour makes it cost-effective for organizations of all sizes, while built-in AWS integrations eliminate complex setup requirements

Service Integration

AWS Elastic Disaster Recovery integrates deeply with core AWS infrastructure services including Amazon EC2 for compute resources, Amazon EBS for storage volumes, and Amazon VPC for network configurations. It leverages AWS Systems Manager for post-launch automation, Amazon CloudWatch for monitoring and metrics, AWS CloudTrail for API logging, and Amazon Route 53 for DNS routing during failover:

- The service uses AWS PrivateLink for secure communication through VPC endpoints, integrates with AWS Direct Connect and VPN for private connectivity, and works with AWS Backup for complementary data protection strategies
- It also supports AWS CloudFormation for infrastructure as code deployments

Onboarding and Offboarding

Customers begin by initializing AWS DRS in their target AWS region and configuring default replication and launch settings through the AWS Management Console. The next step involves installing the AWS Replication Agent on source servers, which automatically begins continuous data replication to the staging area:

- Users then configure individual server launch settings, perform initial recovery drills to validate the setup, and establish their disaster recovery procedures
- The service includes comprehensive getting started guides, quick start tutorials, and free technical training resources to facilitate smooth onboarding and proper configuration of recovery objectives

Getting Started Guide: https://docs.aws.amazon.com/drs/latest/userguide/getting-started.html

Data Removal

Data removal during offboarding involves stopping replication on source servers, terminating recovery instances, and deleting source server resources from the DRS console. Customers must clean up staging area resources including replication servers and associated EBS volumes. Point-in-time snapshots are automatically deleted based on retention policies, but customers should verify complete removal of all DRS-related resources including security groups and VPC endpoints to avoid ongoing charges.

Backup/Restore and Disaster Recovery

AWS DRS provides comprehensive disaster recovery capabilities through continuous block-level replication and automated point-in-time snapshots. It maintains recovery points every 10 minutes for recent data, hourly snapshots for 24 hours, and daily snapshots for 7 days. The service enables rapid recovery with RTOs of minutes and RPOs of seconds, supporting full application stack recovery including operating systems, databases, and configurations.

Backup Capabilities

AWS DRS creates automated point-in-time snapshots as part of its replication process but is primarily a disaster recovery service rather than a backup solution. While it maintains crash-consistent snapshots for recovery purposes, it works complementarily with AWS Backup for comprehensive data protection strategies. Organizations typically use DRS for disaster recovery and AWS Backup for regular backup operations.

Disaster Recovery

AWS Elastic Disaster Recovery is specifically designed as a comprehensive disaster recovery solution, providing the core DR capabilities rather than integrating with other DR services. It serves as the primary disaster recovery platform for protecting on-premises, AWS, and third-party cloud workloads, offering complete failover and failback capabilities across regions and availability zones.

Recovery Objectives

AWS DRS helps customers achieve aggressive recovery objectives with RPOs measured in seconds through continuous block-level replication and RTOs measured in minutes through automated server conversion and launch processes. The service maintains multiple point-in-time snapshots and provides rapid failover capabilities, enabling organizations to meet stringent business continuity requirements while minimizing data loss and downtime during disaster scenarios.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/general/latest/gr/drs.html

FAQ: https://aws.amazon.com/disaster-recovery/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/drs/latest/userguide/what-is-drs.html

User Guide: https://docs.aws.amazon.com/drs/latest/userguide/getting-started.html

API Reference: https://docs.aws.amazon.com/drs/latest/APIReference/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/disaster-recovery/pricing/

Cost Optimization

AWS DRS offers unique cost optimization through its staging area architecture that eliminates expensive idle recovery infrastructure while maintaining continuous protection. The service charges a flat rate of $0.028 per source server per hour, with additional costs only for actual AWS resources used during replication (EBS storage) and recovery operations (EC2 instances):

- Organizations can optimize costs by right-sizing recovery instances, using appropriate EBS volume types, conducting efficient drill testing schedules, and properly cleaning up resources after recovery operations
- The pay-per-protected-server model provides predictable disaster recovery costs without large upfront investments

Savings Considerations

Primary cost savings come from eliminating traditional disaster recovery site infrastructure, maintenance, and staff costs associated with maintaining secondary data centers. The staging area design reduces ongoing costs by using minimal compute resources during normal operations, only scaling up during actual recovery events:

- Organizations save significantly on hardware, facility, and operational expenses while benefiting from AWS's economies of scale
- Additional savings result from reduced recovery testing costs, eliminated hardware refresh cycles, and the ability to use recovered instances permanently in AWS, potentially reducing primary infrastructure costs through hybrid cloud strategies

Service Name

AWS Elemental MediaConnect

Service Overview

AWS Elemental MediaConnect is a reliable, secure, and flexible transport service for live video that enables broadcasters and premium video providers to ingest live video into the AWS Cloud and distribute it to multiple destinations, both within and outside AWS. The service provides the reliability, security, and visibility of traditional distribution methods with the flexibility and cost-effectiveness of internet-based transmission. MediaConnect adds a quality-of-service layer over standard IP transport, enabling uninterrupted live video connections through packet recovery and supports industry-standard video encryption. Users create flows to establish transport between sources and one or more outputs, enabling content sharing through entitlements with other AWS accounts.

Key Use Cases

- Contribution: Ingest content from on-premises encoders into AWS Cloud using transport stream or CDI flows
- Distribution: Deliver content to different geographical locations using multi-region flows
- Entitlements: Share content securely with other AWS accounts through permission-based access
- Replication and Monitoring: Distribute video to multiple destinations and enable real-time monitoring of multiple video signals

Features

- **Multiple Protocol Support**: Supports Zixi, RIST, SRT, RTP, RTP-FEC, NDI, and other industry-standard protocols for reliable video contribution
- **CDI and JPEG XS**: Provides uncompressed and visually-lossless video transport using AWS Cloud Digital Interface and JPEG XS compression
- **Source Failover**: Offers Merge and Failover modes using two redundant sources to minimize disruption
- **End-to-End Encryption**: Built-in security with AES encryption, SPEKE key management, and IP allow listing
- **Real-Time Monitoring**: Provides network health metrics, content quality monitoring, and EventBridge integration
- **Workflow Monitor**: Discover, visualize, and monitor live video resources across workflows
- **Router Integration**: Real-time routing of live video streams between sources and destinations in the cloud

Value Proposition

MediaConnect offers significant advantages over traditional satellite and fiber solutions including faster deployment (minutes vs months), pay-as-you-go pricing without upfront commitments, and integrated AWS billing and support. Unlike infrastructure providers, MediaConnect provides better long-distance performance, supports multiple protocols, enables both contribution and business-to-business distribution workflows, and includes built-in security and sharing capabilities. The service delivers professional-grade quality of service, strong encryption, flexible routing, comprehensive monitoring, and seamless integration with other AWS Media Services, making it ideal for broadcasters who need reliable, secure, and cost-effective live video transport.

Service Integration

MediaConnect integrates seamlessly with AWS Media Services ecosystem including AWS Elemental MediaLive for live video processing and encoding, AWS Elemental MediaPackage for content packaging and distribution, AWS Elemental MediaStore for media storage, and AWS Elemental MediaTailor for ad insertion. The service also integrates with Amazon EventBridge for real-time event monitoring, AWS CloudWatch for metrics and alarms, AWS Secrets Manager for SPEKE encryption key management, Amazon VPC for secure network connectivity, AWS Direct Connect for dedicated network connections, Amazon S3 for backup storage, and Amazon Route 53 Application Recovery Controller for disaster recovery orchestration.

Onboarding and Offboarding

Customers get started by creating an AWS account and setting up IAM users with necessary permissions. The initial setup involves accessing the MediaConnect console, creating a flow with specified source details including transport protocol and encryption type, and configuring destinations:

- Users can optionally store encryption keys in AWS Secrets Manager and grant MediaConnect permission to access them
- The service provides step-by-step tutorials for common scenarios like creating flows for contribution, distribution, and entitlements
- Getting started requires minimal infrastructure setup as MediaConnect is a fully managed service that can be deployed in minutes using the AWS Management Console, CLI, or API

Getting Started Guide: https://docs.aws.amazon.com/mediaconnect/latest/ug/getting-started.html

Data Removal

When offboarding from MediaConnect, customers must stop and delete all flows to avoid continued charges. Data removal involves deleting flow configurations, source settings,

output destinations, and entitlements through the console, CLI, or API. Since MediaConnect is a transport service for live video streams, it does not persistently store customer video content, but customers should ensure all associated resources like encryption keys in Secrets Manager and related monitoring configurations are properly cleaned up.

Backup/Restore and Disaster Recovery

MediaConnect supports source failover with redundant sources using Merge and Failover modes to minimize service disruption. The service provides built-in reliability through redundant infrastructure and automatic recovery from issues. For disaster recovery, customers can deploy flows across multiple AWS regions and use Amazon Route 53 Application Recovery Controller for automated failover orchestration. However, as a live video transport service, MediaConnect does not provide traditional backup capabilities for video content itself.

Backup Capabilities

MediaConnect does not provide traditional backup capabilities as it is a live video transport service that does not persistently store video content. The service focuses on real-time transport and replication of live streams. Backup strategies would need to be implemented at the source or destination endpoints, such as using AWS Elemental MediaLive to create archive outputs to Amazon S3 for content backup purposes.

Disaster Recovery

MediaConnect supports disaster recovery through multi-region deployment capabilities, source failover with redundant sources, and integration with Amazon Route 53 Application Recovery Controller for automated failover. The service provides built-in infrastructure redundancy and automatic recovery mechanisms. However, it does not integrate with AWS Elastic Disaster Recovery as MediaConnect is a live streaming transport service rather than a traditional application or data workload.

Recovery Objectives

MediaConnect helps customers achieve low RPO and RTO objectives through source failover capabilities that can switch between redundant sources in 500 milliseconds during failures. Multi-region deployment enables quick regional failover for distribution flows. The service's built-in redundancy and automatic recovery mechanisms minimize downtime. For live video workflows, customers can achieve near-zero RPO through real-time replication and sub-second RTO through automated failover mechanisms.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/mediaconnect/latest/ug/quotas.html

FAQ: https://aws.amazon.com/mediaconnect/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/mediaconnect/latest/ug/what-is.html

User Guide: https://docs.aws.amazon.com/mediaconnect/latest/ug/what-is.html

API Reference: https://docs.aws.amazon.com/mediaconnect/latest/api/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/mediaconnect/pricing/

Cost Optimization

MediaConnect offers reserved outbound bandwidth pricing with up to 12-month commitments for discounted internet data transfer rates. The service uses pay-as-you-go pricing with no upfront commitments, charging only for running flows and data transfer. Cost optimization strategies include stopping flows when not in use (idle flows incur no charges), using reserved bandwidth for predictable workloads, optimizing bitrates and compression settings, leveraging VPC outputs to avoid internet transfer charges, and selecting appropriate flow types (transport stream vs CDI) based on quality requirements to balance cost and performance.

Savings Considerations

Main cost savings come from eliminating traditional satellite and fiber infrastructure with their high upfront costs and long-term commitments. MediaConnect's pay-as-you-go model allows customers to scale dynamically without overprovisioning:

- Reserved outbound bandwidth pricing provides additional savings for consistent usage patterns
- The service reduces operational costs through automated management, integrated billing with other AWS services, and elimination of hardware maintenance
- Regional data transfer within AWS avoids internet egress charges, and the ability to start/stop flows on-demand prevents unnecessary charges during inactive periods, making it significantly more cost-effective than traditional broadcast infrastructure

Service Name

AWS Elemental MediaConvert

Service Overview

AWS Elemental MediaConvert is a file-based video processing service designed for content owners and distributors with media libraries of any size. It provides scalable video transcoding that converts input video files into output files with various formats and resolutions for delivery to multiple devices. The service offers broadcast-grade features including professional codecs supporting increased bit depth and HDR content, static graphic overlays, advanced audio processing, digital rights management (DRM), and closed captioning support. MediaConvert enables premium content experiences through adaptive bitrate (ABR) packaging output formats for delivering high-quality content from various sources to primary and multiscreen devices.

Key Use Cases

- Video-on-demand content creation: Converting media files for broadcast and multiscreen delivery platforms with adaptive streaming capabilities
- Content format conversion: Transcoding media files for compatibility, compression, and distribution across different devices and platforms
- Broadcast-grade processing: Creating premium content with professional codecs, HDR support, advanced audio, and closed captioning for television and streaming services

Features

- **Professional Broadcast Codecs**: Support for MPEG-2, AVC, AV1, HEVC with 10-bit 4:2:2 color sampling and HDR content creation
- **Adaptive Bitrate Packaging**: CMAF, HLS, DASH, and MSS output formats with automated ABR configuration
- **Advanced Audio Features**: Multi-language audio, audio loudness normalization, Dolby Atmos support
- **Content Protection**: Digital rights management (DRM) integration and encryption capabilities
- **Graphics and Overlays**: Static graphic overlays, image insertion, motion graphics, video overlays
- **Quality Features**: Quality-defined variable bitrate (QVBR), HDR processing including Dolby Vision

Value Proposition

MediaConvert eliminates the complexity of building and operating video processing infrastructure while providing broadcast-grade capabilities that other cloud encoding services often lack. It offers automatic scalability with elastic resource provisioning, pay-as-you-go pricing with tiered volume discounts, and built-in redundancy across multiple Availability Zones:

- The service provides comprehensive broadcast-quality features including static graphic overlays, audio loudness normalization, SCTE-35 ad insertion support, DRM integration, and extensive closed captioning capabilities
- Customers benefit from short turnaround times even with varying workloads, integrated billing and monitoring with other AWS services, and access to the latest codecs and processing technologies without infrastructure management overhead

Service Integration

MediaConvert integrates seamlessly with Amazon S3 for input and output file storage, Amazon CloudFront for global content delivery and CDN services, and AWS Lambda for creating automated watch folders and workflow orchestration. It works with Amazon EventBridge for job status notifications and workflow automation, AWS Step Functions for complex video processing workflows, and Amazon CloudWatch for monitoring and logging encoding metrics:

- The service supports IAM for access control and permissions management, integrates with AWS Systems Manager for resource monitoring, and can be used with Amazon SNS for notifications
- MediaConvert also works with AWS Elemental MediaPackage for content preparation and delivery, and supports integration with third-party DRM providers for content protection

Onboarding and Offboarding

Customers get started by ensuring their input files are accessible in Amazon S3, HTTP, or HTTPS locations and setting up appropriate IAM permissions for MediaConvert to access these resources. They can create their first transcoding job through the MediaConvert console by specifying input files and output settings, or use the API/SDK for programmatic access:

- The console provides system presets for common outputs, while advanced users can create custom presets and job templates
- MediaConvert requires jobs and file storage to be in the same AWS Region
- Customers can track job status through EventBridge notifications and manage resources using queues for parallel processing optimization

Getting Started Guide: https://docs.aws.amazon.com/mediaconvert/latest/ug/getting-started.html

Data Removal

MediaConvert processes files stored in customer-controlled Amazon S3 buckets, so data removal involves deleting input and output files from S3 storage locations. Customers retain full control over their media files and can delete them at any time to avoid ongoing storage charges. The service itself does not store customer content permanently - it only processes files during transcoding jobs and outputs results to specified S3 locations.

Backup/Restore and Disaster Recovery

MediaConvert runs on redundant infrastructure distributed across physically separated Availability Zones with automatic health monitoring and component replacement without job disruption. The service provides built-in redundancy and fault tolerance, automatically replacing degraded components. However, backup of source media files and outputs relies on customer-managed Amazon S3 storage with cross-region replication and versioning capabilities.

Backup Capabilities

MediaConvert does not provide native backup capabilities as it is a processing service. Backup depends on underlying Amazon S3 storage where customers store input and output files. Customers can enable S3 versioning, cross-region replication, and use AWS Backup for S3 bucket protection to ensure media file availability.

Disaster Recovery

MediaConvert provides high availability through multi-AZ infrastructure with automatic failover, but does not directly integrate with AWS Elastic Disaster Recovery. Disaster recovery relies on S3's cross-region replication for source files and the ability to recreate transcoding jobs in different regions using job templates and presets.

Recovery Objectives

MediaConvert helps achieve low RTO through its multi-AZ redundant infrastructure that automatically recovers from component failures without job disruption. For RPO objectives, customers rely on S3's versioning and cross-region replication capabilities for source media protection, enabling rapid job recreation in alternate regions when needed.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/general/latest/gr/mediaconvert.html

FAQ: https://aws.amazon.com/mediaconvert/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/mediaconvert/latest/ug/what-is.html

User Guide: https://docs.aws.amazon.com/mediaconvert/latest/ug/

API Reference: https://docs.aws.amazon.com/mediaconvert/latest/apireference/

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/mediaconvert/pricing/

Cost Optimization

MediaConvert offers tiered volume-based pricing that automatically provides discounts as usage increases monthly, with savings up to 19% for large archives. The service supports Quality-Defined Variable Bitrate (QVBR) encoding to reduce file sizes while maintaining quality, and HEVC encoding for significant file size reduction:

- Reserved pricing is available for predictable workloads with 12-month commitments offering fixed monthly rates
- Accelerated transcoding completes jobs up to 25 times faster for time-sensitive workflows
- Two-tier pricing structure allows choosing Basic tier for simple web distribution or Professional tier only when advanced features are needed

Savings Considerations

Primary cost savings come from eliminating on-premises infrastructure investments and maintenance costs, with pay-as-you-go pricing ensuring customers only pay for actual usage. Automated tiered pricing provides volume discounts without upfront commitments, while file size reduction through QVBR and HEVC encoding can save substantial storage and CDN delivery costs:

- Large content producers can save 18.5% on monthly transcoding through volume discounts
- Organizations with 300 TB archives can save over $140,000 annually through optimized transcoding and file size reduction
- The elastic scaling eliminates need for capacity planning and reduces turnaround times during peak workloads

Service Name

AWS Elemental MediaLive

Service Overview

AWS Elemental MediaLive is a broadcast-grade live video processing service that enables video providers to encode high-quality live video streams for delivery to broadcast televisions and multiscreen devices. The service transforms live video content from one format and package into other formats and packages to provide compatibility with various playback devices such as smartphones and set-top boxes. MediaLive supports industry-standard video codecs including H.264/AVC and H.265/HEVC, along with communication protocols like RTP, HLS, and RTSP. It operates on redundant infrastructure distributed across Availability Zones and automatically replaces degraded components without disrupting channels.

Key Use Cases

- Live streaming: Processing real-time live video into adaptive bitrate streams for delivery to various devices and platforms
- Broadcast television: Encoding high-quality live video streams with broadcast capabilities like ad markers, closed captions, and multiple language audio tracks
- 24x7 linear workflows: Building flexible continuous streaming workflows for channels that run around the clock
- Event-based live streams: Creating live streams for specific events with dynamic input switching and scheduling capabilities
- Multi-program transport streams (MPTS): Creating UDP transport streams carrying multiple programs for broadcast distribution

Features

- **Real-time Video Encoding**: Processes live video in real-time and converts it into adaptive bitrate streams with support for H.264/AVC, H.265/HEVC, and AV1 codecs
- **Broadcast Capabilities**: Supports ad markers, closed captions, multiple language audio tracks, audio descriptors, and FCC-mandated loudness correction
- **High Availability**: Runs on redundant infrastructure across Availability Zones with automatic failover and component replacement without channel disruption
- **Statistical Multiplexing**: Optimizes bandwidth usage by combining multiple audio and video streams into a single output stream for broadcast distribution
- **Input Switching**: Supports multiple inputs per channel with scheduled switching between different video sources
- **Motion Graphics**: Enables overlay of graphics, logos, and other visual elements on live video streams

- **Live Video Workflow Monitor**: Provides monitoring and alerting capabilities for live video processing workflows
- **Automated Resource Provisioning**: Automatically deploys, scales, and monitors encoding resources based on demand

Value Proposition

AWS Elemental MediaLive offers significant advantages over alternatives through its broadcast-grade quality with 99.9% SLA, fully managed infrastructure that eliminates hardware management overhead, and seamless integration with other AWS Media Services. The service provides automatic scaling and redundancy across Availability Zones, ensuring high availability without manual intervention:

- Customers benefit from flexible pricing models including on-demand and reserved instances with up to 70% savings, comprehensive support for industry standards, and the ability to build both 24x7 linear workflows and event-based streaming
- The service's cloud-native architecture enables rapid deployment and global reach while maintaining broadcast-quality output

Service Integration

MediaLive integrates extensively with AWS Media Services ecosystem including AWS Elemental MediaPackage for video packaging and origination, AWS Elemental MediaConnect for high-quality video transport, AWS Elemental MediaStore for media-optimized storage, and AWS Elemental MediaTailor for ad insertion and monetization. Core AWS integrations include Amazon CloudWatch for monitoring and metrics, AWS Identity and Access Management (IAM) for security, Amazon S3 for asset storage and output destinations, AWS Systems Manager for password management, Amazon VPC for private networking, and Amazon EventBridge for event-driven workflows. The service also works with AWS Resource Groups for tagging and organization.

Onboarding and Offboarding

Customers can get started with MediaLive through three approaches: the workflow wizard for simplified setup supporting common input/output types, following the tutorial for console introduction with additional features, or designing workflows from scratch for advanced customization. Prerequisites include an AWS account with appropriate IAM permissions and completion of preliminary setup steps:

- The workflow wizard handles the entire MediaLive workflow including creating channels, inputs, and related resources in other AWS services like MediaPackage
- Customers can begin with basic configurations and progressively add features like input switching, motion graphics, and advanced encoding settings

Getting Started Guide: https://docs.aws.amazon.com/medialive/latest/ug/getting-started.html

Data Removal

MediaLive follows the AWS shared responsibility model where customers are responsible for managing their content and security configurations. When offboarding, customers must delete their channels, inputs, and associated resources through the MediaLive console or API. Content stored in integrated services like S3 or MediaStore must be deleted separately. MediaLive does not automatically retain customer data after resource deletion, and customers should follow organizational best practices for handling sensitive data during the offboarding process.

Backup/Restore and Disaster Recovery

MediaLive provides built-in disaster recovery through its distributed architecture across multiple Availability Zones with automatic failover capabilities. The service does not require traditional backup as it processes live streams in real-time. For configuration backup, customers can export channel configurations and use Infrastructure as Code tools like CloudFormation. Content backup depends on integrated services - outputs sent to S3 can use S3 versioning and cross-region replication, while MediaPackage provides its own redundancy features.

Backup Capabilities

MediaLive does not have traditional backup capabilities as it is a real-time video processing service. However, it provides configuration export capabilities and supports Infrastructure as Code for recreating channels:

- The service does not integrate directly with AWS Backup since it processes live streams rather than storing persistent data
- Customers rely on the service's high availability features and can backup channel configurations manually or through automation

Disaster Recovery

MediaLive supports disaster recovery through its multi-Availability Zone architecture with automatic failover and component replacement without service interruption. The service does not integrate with AWS Elastic Disaster Recovery as it is designed for real-time processing with built-in resilience. Customers can implement cross-region disaster recovery by replicating channel configurations and using multiple regions for critical workflows, though this requires manual setup and management.

Recovery Objectives

MediaLive helps customers achieve low RTO objectives through its automatic failover mechanisms across Availability Zones, typically recovering from failures within minutes without manual intervention. For RPO, since MediaLive processes live streams in real-time, data loss is minimized to the duration of the failover process. The service's redundant

pipeline architecture ensures continuous operation, and customers can implement multi-region deployments for even lower RTO/RPO requirements for critical broadcast workflows.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/general/latest/gr/medialive_region.html

FAQ: https://aws.amazon.com/medialive/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/medialive/latest/ug/what-is.html

User Guide: https://docs.aws.amazon.com/medialive/latest/ug/what-is.html

API Reference: https://docs.aws.amazon.com/medialive/latest/apireference/what-is.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/medialive/pricing/

Cost Optimization

MediaLive offers reserved pricing with up to 70% savings for channels running more than 180 hours per month with 12-month commitments. Cost optimization strategies include using single-pipeline channels for non-critical workloads versus standard dual-pipeline for production, optimizing input/output configurations to match actual requirements, and leveraging statistical multiplexing for broadcast distribution efficiency:

- Customers can implement data-driven reservation management using AWS Cost and Usage Reports with Amazon QuickSight for utilization analysis
- Idle resource management is important as push inputs incur costs when idle, while pull inputs do not

Savings Considerations

Primary cost savings come from reserved instances offering up to 70% discount over on-demand pricing for consistent workloads. Customers should analyze usage patterns using Cost Explorer and implement optimal balance between reserved and on-demand instances:

- Data transfer costs can be minimized by keeping MediaLive outputs within AWS or delivering to CloudFront
- Statistical multiplexing provides bandwidth optimization for broadcast scenarios

- Regular monitoring of reservation utilization through QuickSight dashboards and Cost and Usage Reports ensures continued cost effectiveness
- Proper resource sizing and avoiding over-provisioning of encoding capabilities drives additional savings

Service Name

AWS Elemental MediaPackage

Service Overview

AWS Elemental MediaPackage is a cloud-based video packaging and origination service that enables the delivery of secure, scalable, and reliable video streams to various playback devices and content delivery networks (CDNs). The service provides just-in-time packaging that dynamically customizes live video streams and VOD assets for delivery in formats compatible with requesting devices. It simplifies video workflows by converting incoming content on-the-fly from a single format to multiple delivery formats while applying Digital Rights Management (DRM) standards, enabling broadcast-quality video content preparation and protection for multiscreen delivery.

Key Use Cases

- Live streaming for broadcast networks and sports leagues: Real-time packaging and delivery of live video content with adaptive bitrate streaming and DRM protection
- Video-on-demand content delivery: Packaging and distribution of pre-recorded video content with multiple format outputs and content protection
- Over-the-top (OTT) video services: Secure multiscreen content delivery to various devices including tablets, smartphones, connected TVs, and set-top boxes with time-shifted viewing capabilities

Features

- **Just-in-Time Packaging**: Dynamically packages content on-demand rather than pre-processing, reducing storage requirements and enabling efficient bandwidth usage
- **Multi-Format Output Support**: Supports MPEG-DASH, HLS, Microsoft Smooth Streaming, and CMAF packaging formats for broad device compatibility
- **Digital Rights Management**: Integrates with multiple DRM technologies including Google Widevine, Microsoft PlayReady, and Apple FairPlay for content protection
- **Adaptive Bitrate Streaming**: Automatically adjusts video quality based on viewer's network conditions for optimal viewing experience
- **Time-Shifted Viewing**: Enables DVR-like features including start-over, catch-up TV, and time delay for live content
- **Input Redundancy**: Creates two ingest URLs for automatic failover if primary input fails, ensuring uninterrupted delivery
- **CDN Authorization**: Supports secure content delivery through Amazon CloudFront and other CDNs with header-based authorization

- **Live-to-VOD Asset Creation**: Extracts VOD assets from live streams through harvest jobs for later on-demand viewing

Value Proposition

AWS Elemental MediaPackage provides superior value through its highly scalable, managed service approach that eliminates the complexity of implementing and managing packaging infrastructure. Unlike third-party solutions running on virtual servers, MediaPackage operates as a native cloud service with embedded redundancy, automatic scaling, and separation of ingest from egress for greater scalability:

- The service offers more control and customization options compared to CDN-based packaging services, supports multi-CDN strategies, and provides CDN-agnostic flexibility
- Key advantages include broadcast-grade reliability with distributed resources across multiple Availability Zones, seamless integration with other AWS Media Services, comprehensive DRM support, and cost efficiency through just-in-time packaging that reduces storage requirements while supporting multiple output formats from a single source

Service Integration

MediaPackage integrates deeply with the AWS Media Services ecosystem, particularly AWS Elemental MediaLive for live video encoding and processing, where MediaLive channels can directly output HLS streams to MediaPackage using WebDAV authentication. Amazon CloudFront serves as the preferred CDN for content delivery, offering optimized data transfer rates and built-in authorization support:

- Amazon S3 provides storage for VOD assets and live-to-VOD content through harvest jobs
- AWS Elemental MediaTailor enables targeted ad insertion and monetization
- Amazon CloudWatch monitors service metrics including content input/output counts and response times
- AWS Identity and Access Management (IAM) controls access permissions and authentication
- AWS Secrets Manager stores encryption keys for DRM implementation
- The service also works with AWS Elemental MediaStore for high-performance media storage and can integrate with third-party encoders and CDNs for hybrid deployment strategies

Onboarding and Offboarding

Getting started with MediaPackage involves creating an AWS account with appropriate IAM permissions, then following a structured approach through the console, CLI, or API. The process begins by creating a channel group as the top-level organizational resource,

followed by creating channels that serve as entry points for incoming live content from encoders like AWS Elemental MediaLive:

- Next, customers create endpoints attached to channels that represent content outputs with specific packaging formats and configurations
- For live workflows, upstream encoders are configured with WebDAV to push streams using channel input URLs, usernames, and passwords
- For VOD workflows, customers create packaging groups and configurations, then create assets that ingest source content from Amazon S3
- The service provides immediate access to playback URLs for downstream CDNs and players once resources are created

Getting Started Guide:
https://docs.aws.amazon.com/mediapackage/latest/userguide/getting-started.html

Data Removal

Offboarding requires systematic deletion of MediaPackage resources in hierarchical order: endpoints must be deleted before channels, channels before channel groups, and assets before packaging configurations and groups. All content stored temporarily for time-shifted viewing is automatically purged based on retention policies. For persistent data like live-to-VOD assets stored in Amazon S3, customers can use List and Get APIs to copy data to alternative storage before deletion. The console, CLI, and API provide deletion capabilities with confirmation dialogs to prevent accidental removal.

Backup/Restore and Disaster Recovery

MediaPackage provides built-in disaster recovery through its high-availability architecture distributed across multiple AWS Availability Zones within regions. The service includes automatic input redundancy with dual ingest URLs for seamless failover if primary inputs fail. Content is temporarily stored based on configurable time-shifted viewing windows (up to 14 days) but this is not designed for backup purposes. For persistent data protection, live-to-VOD harvest jobs can store content in Amazon S3 with standard S3 durability and backup features. Cross-region failover capabilities enable content delivery from alternative regions during outages.

Backup Capabilities

MediaPackage does not provide traditional backup capabilities as it focuses on real-time content packaging and delivery. The service temporarily stores content for time-shifted viewing (maximum 14 days) but this serves operational purposes rather than backup:

- Persistent content protection relies on Amazon S3 integration through live-to-VOD harvest jobs, which can leverage S3's backup features and AWS Backup service
- The service does not directly integrate with AWS Backup for its own operational data

Disaster Recovery

MediaPackage supports disaster recovery through high-availability architecture across multiple Availability Zones and cross-region failover capabilities. The service automatically scales and manages infrastructure failures with built-in redundancy:

- However, it does not integrate with AWS Elastic Disaster Recovery as it operates as a managed service with its own resilience mechanisms
- Recovery strategies focus on service-level failover and content delivery continuity rather than traditional disaster recovery approaches

Recovery Objectives

MediaPackage helps achieve low Recovery Point Objectives (RPO) through real-time content processing and automatic input redundancy that prevents data loss during failover. Recovery Time Objectives (RTO) are optimized through instant endpoint activation, automatic scaling, and cross-region failover capabilities. The service's distributed architecture and managed infrastructure minimize downtime during disruptions. For persistent content, live-to-VOD harvest jobs combined with S3 storage provide additional RPO protection with configurable retention policies.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/mediapackage/latest/userguide/quotas.html

FAQ: https://aws.amazon.com/mediapackage/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/mediapackage/latest/userguide/what-is.html

User Guide: https://docs.aws.amazon.com/mediapackage/latest/userguide/getting-started.html

API Reference: https://docs.aws.amazon.com/mediapackage/latest/apireference/what-is.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/mediapackage/pricing/

Cost Optimization

MediaPackage offers several cost optimization opportunities through its pay-as-you-go model with charges based on content ingestion and origination volumes measured in GB. Just-in-time packaging reduces storage costs by requiring only single-format source content that gets packaged into multiple formats on-demand:

- Using Amazon CloudFront as the CDN significantly reduces origination costs through high cache hit ratios (typically 99% for live content, 80% for VOD), as cached content doesn't generate additional MediaPackage charges
- Customers can optimize by rightsizing input streams, implementing efficient encoder settings, and leveraging adaptive bitrate streaming
- For high-volume customers (typically 10TB+ monthly), annual commitment discounts are available
- The service eliminates infrastructure management costs and provides automatic scaling without over-provisioning

Savings Considerations

Primary cost savings come from eliminating the need for separate packaging infrastructure and reducing storage requirements through single-format source content strategy. Using Amazon CloudFront optimizes costs through reduced data transfer charges compared to third-party CDNs and improved cache efficiency:

- The managed service model eliminates operational overhead, monitoring costs, and manual scaling requirements
- Input redundancy is built-in without additional charges, reducing failover infrastructure costs
- Multi-format output from single input reduces encoding and storage costs compared to pre-processing multiple formats
- VOD packaging efficiency allows significant storage reduction as only one adaptive bitrate set is required rather than multiple pre-packaged formats
- Automatic scaling prevents over-provisioning costs while ensuring performance during peak demand
- Regional deployment options help optimize data transfer costs based on audience geography

Service Name

AWS Elemental MediaTailor

Service Overview

AWS Elemental MediaTailor is a scalable ad insertion and channel assembly service that enables targeted advertisement content delivery and creation of linear streaming channels for over-the-top (OTT) video applications while maintaining broadcast quality. The service operates as a manifest-only solution, supporting server-side ad insertion (SSAI) for both live and video-on-demand (VOD) content across HTTP Live Streaming (HLS) and MPEG Dynamic Adaptive Streaming over HTTP (DASH) protocols. MediaTailor generates unique manifest files for each viewer, delivers personalized ads, and assembles linear channels using existing VOD and live content without touching the actual content segments.

Key Use Cases

- Server-side ad insertion: Insert personalized, targeted advertisements into live and VOD streams while maintaining broadcast quality and seamless playback
- Channel assembly: Create linear streaming channels using existing VOD content and live sources to build traditional broadcast-style experiences
- Sports and live event monetization: Deliver targeted ads during live sports broadcasts and events with precise timing and minimal latency
- OTT platform monetization: Enable streaming services to monetize their content libraries through personalized ad experiences
- Multi-platform ad delivery: Support cross-device ad insertion for web browsers, mobile apps, smart TVs, and set-top boxes

Features

- **Server-Side Ad Insertion (SSAI)**: Inserts advertisements upstream before delivery, ensuring seamless playback without discrimination between content and ads
- **Channel Assembly**: Creates linear streaming channels using existing VOD content and live sources without touching content segments
- **Dynamic Transcoding**: Automatically matches ad quality and format to primary video content for optimal viewing experience
- **HLS Interstitials Support**: Enables insertion of interstitial advertisements directly into live streams using HLS Interstitials specification
- **Advanced Ad Tracking**: Provides server-side beaconing, beacon deduplication, adaptive throttling, and in-order beacon delivery
- **Time-Shifted Viewing**: Supports DVR-like functions including start-over, pause, rewind, and fast-forward capabilities

- **Multi-Protocol Support**: Supports HLS, DASH, and CMAF streaming protocols for broad device compatibility
- **Personalization**: Delivers targeted ads based on viewer demographics, interests, and location preferences

Value Proposition

MediaTailor offers superior server-side ad insertion that maintains broadcast-quality streaming while providing personalized ad experiences. Unlike client-side solutions, it eliminates ad blockers' ability to skip ads and ensures consistent quality across all devices:

- The service automatically scales with viewership, requires no upfront commitments or contracts, and integrates seamlessly with existing AWS Media Services workflows
- Key advantages include frame-accurate ad transitions, comprehensive measurement and reporting capabilities, support for industry standards like VAST/VMAP/VPAID, and the ability to monetize both live and VOD content effectively
- MediaTailor reduces operational complexity by handling ad transcoding, manifest manipulation, and beacon tracking automatically

Service Integration

MediaTailor integrates deeply with AWS Media Services ecosystem, particularly AWS Elemental MediaPackage for content packaging and distribution, AWS Elemental MediaLive for live video processing, and AWS Elemental MediaConnect for secure video transport. The service leverages Amazon CloudFront as a CDN for global content delivery and manifest distribution:

- Integration with AWS Identity and Access Management (IAM) provides secure access control, while Amazon CloudWatch enables comprehensive monitoring and logging
- MediaTailor also connects with AWS Elemental MediaConvert for on-demand transcoding needs and can utilize Amazon S3 for content storage and Amazon Data Firehose for log delivery

Onboarding and Offboarding

Customers begin by accessing the MediaTailor console through their AWS account and creating a playback configuration with required settings including a unique name, content source URL prefix, and ad decision server URL. The service offers two main tutorials: ad insertion for incorporating personalized ads into content streams, and channel assembly for creating linear streaming channels:

- Getting started requires preparing HLS or DASH compatible manifests, configuring ad decision server integration, and setting up optional features like CDN integration, logging, and personalization thresholds
- MediaTailor provides comprehensive documentation, API references, and step-by-step guides to facilitate smooth onboarding

Getting Started Guide: https://docs.aws.amazon.com/mediatailor/latest/ug/getting-started.html

Data Removal

Data removal involves deleting MediaTailor configurations, channels, and associated resources through the console, CLI, or API. Customers should clean up playback configurations, channel assembly components including channels, source locations, and VOD sources. Log data stored in Amazon CloudWatch, S3, or Data Firehose follows respective service retention policies. Since MediaTailor operates as a manifest manipulation service without storing actual video content, data removal primarily involves configuration and metadata cleanup rather than content deletion.

Backup/Restore and Disaster Recovery

MediaTailor leverages AWS global infrastructure built on Regions and Availability Zones for inherent resilience and disaster recovery. The service automatically handles failover between Availability Zones without interruption. Configuration data and channel assembly settings are maintained within AWS managed infrastructure. Customers can export configuration settings programmatically and recreate them in different regions if needed. The service's manifest-only architecture minimizes data loss risks since actual video content remains in customer-controlled storage locations.

Backup Capabilities

MediaTailor does not provide traditional backup capabilities as it operates as a manifest manipulation service without storing video content. Configuration data is maintained within AWS managed infrastructure with built-in redundancy:

- Customers can programmatically export and recreate configurations using APIs or Infrastructure as Code tools
- The service does not directly integrate with AWS Backup since it primarily manages metadata and configurations rather than data requiring traditional backup operations

Disaster Recovery

MediaTailor supports disaster recovery through AWS's multi-AZ infrastructure and regional deployment options. The service automatically provides failover capabilities within regions through Availability Zone redundancy:

- For cross-region disaster recovery, customers can recreate configurations and channel assembly setups in alternate regions
- MediaTailor does not integrate with AWS Elastic Disaster Recovery as it manages configuration metadata rather than application workloads requiring replication and failover orchestration

Recovery Objectives

MediaTailor helps achieve low RTO objectives through its managed service architecture that automatically handles infrastructure failover within AWS Availability Zones. RPO is minimal since the service primarily manages configuration data that can be quickly recreated. Customers can implement cross-region setups for additional resilience, with configuration recreation possible within minutes using APIs or Infrastructure as Code. The manifest-only architecture ensures that actual content remains available from origin sources during recovery scenarios.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/mediatailor/latest/ug/quotas.html

FAQ: https://aws.amazon.com/mediatailor/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/mediatailor/latest/ug/what-is.html

User Guide: https://docs.aws.amazon.com/mediatailor/latest/ug/

API Reference: https://docs.aws.amazon.com/mediatailor/latest/apireference/

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/mediatailor/pricing/

Cost Optimization

MediaTailor offers unique cost optimization through its new pricing model with VOD ad insertions priced at 50% of live streaming costs ($0.25 vs $0.50 per 1,000 insertions). The service includes the first 10 ad transcodes per 1,000 insertions at no charge, with additional transcoding charged at MediaConvert rates:

- Channel assembly utilizes existing encoded content without re-transcoding, minimizing processing costs

- Features like MediaTailor Vended Logs and Log Filtering reduce logging expenses by allowing selective log delivery to cost-effective storage options like S3 instead of CloudWatch
- CDN integration optimization and cache configuration can significantly reduce origin request costs

Savings Considerations

Primary savings come from the 50% reduced VOD pricing model compared to live streams, making ad-supported VOD libraries more cost-effective. Included ad transcoding allowances (first 10 per 1,000 insertions) reduce processing costs for standard use cases:

- Channel assembly leverages existing content assets without additional transcoding, maximizing ROI on content libraries
- Strategic CDN configuration with proper caching reduces origin requests and data transfer costs
- Log filtering and alternative storage options (S3 vs CloudWatch) can reduce operational expenses by up to 90% for logging-intensive deployments
- Volume commitments above 60 million monthly insertions qualify for additional discounted pricing tiers

Service Name

AWS Entity Resolution

Service Overview

AWS Entity Resolution is a fully-managed service that helps organizations match, link, and enhance related records stored across multiple applications, channels, and data stores. The service enables users to create unified views of customer interactions, product information, and business data by using advanced matching techniques including rule-based matching, machine learning-based matching, and data service provider-led matching. It processes records containing identifiers such as names, email addresses, phone numbers, and addresses to accurately link related information without requiring expertise in entity resolution or machine learning. The service supports both batch processing for large datasets and near real-time matching for time-sensitive use cases.

Key Use Cases

- Customer Identity Resolution: Creating unified customer profiles by matching records across CRM systems, marketing platforms, and customer data platforms to eliminate duplicate identities and fragmented customer views
- Healthcare Patient Matching: Building unified patient indexes by linking patient records across multiple healthcare systems and providers to ensure accurate patient identification and comprehensive medical histories
- Product Catalog Management: Matching and linking product information across different systems using various codes (SKU, UPC) to maintain consistent product catalogs and inventory management
- Fraud Detection and Financial Services: Identifying duplicate accounts, verifying customer identities, and detecting fraudulent activities by matching customer records across financial institutions and transaction systems
- Marketing and Advertising Optimization: Enhancing customer targeting and personalization by linking consumer interactions across multiple touchpoints and channels for more effective cross-channel campaigns

Features

- **Flexible Data Preparation**: Supports up to 20 data inputs from AWS Glue with customizable schema mapping, data normalization capabilities, and encryption support for secure data processing
- **Rule-Based Matching**: Ready-to-use deterministic matching rules with advanced fuzzy matching algorithms including Levenshtein Distance, Cosine Similarity, and Soundex for handling data variations and typos

- **Machine Learning Matching**: Pre-configured ML models that analyze entire records together, handle missing fields, and provide confidence scores for match quality assessment
- **Data Service Provider Integration**: Integration with licensed datasets and third-party data service providers for data enrichment and enhanced matching capabilities
- **Near Real-Time Matching**: GetMatchId and GenerateMatchId APIs for synchronous lookup and matching of entities within seconds for time-sensitive applications
- **Automated Processing**: Support for both manual bulk processing and automatic incremental processing to keep matches updated as new data arrives in S3
- **Data Protection**: Built-in encryption with AWS KMS support, data regionalization, and secure data handling with hashing capabilities for PII protection

Value Proposition

AWS Entity Resolution provides a fully-managed, no-code solution that eliminates the complexity of building custom entity matching systems. Organizations choose this service because it requires no expertise in entity resolution or machine learning, offers flexible matching approaches from deterministic rules to advanced ML models, and integrates seamlessly with existing AWS data infrastructure:

- The service delivers significant time-to-value by providing pre-configured matching techniques while maintaining the flexibility to customize rules and workflows
- Key differentiators include advanced fuzzy matching algorithms for handling real-world data imperfections, near real-time matching capabilities for time-sensitive applications, built-in security and compliance features, and the ability to process datasets ranging from thousands to billions of records with automatic scaling

Service Integration

AWS Entity Resolution integrates deeply with core AWS data and analytics services to provide a comprehensive data processing ecosystem. Primary integrations include Amazon S3 for data input and output storage, AWS Glue for data cataloging and schema management, and AWS Glue Data Catalog for metadata management:

- Security integrations include AWS KMS for encryption key management, AWS IAM for access control, and AWS CloudTrail for audit logging
- Monitoring and observability are supported through Amazon CloudWatch for metrics and logging, Amazon CloudWatch Logs for workflow monitoring, and Amazon Data Firehose for log delivery
- The service also integrates with Amazon Connect Customer Profiles for unified customer views, AWS Step Functions for workflow orchestration, AWS Lambda for custom processing, and Amazon EventBridge for event-driven architectures

Onboarding and Offboarding

Getting started with AWS Entity Resolution follows a straightforward process. Customers first prepare their data in Amazon S3 and use AWS Glue crawlers to create data catalog entries:

- Next, they create schema mappings in the Entity Resolution console to define how their data fields should be interpreted for matching
- The third step involves creating matching workflows by selecting matching techniques (rule-based, ML-based, or provider service-based) and configuring output destinations
- Customers can then run workflows either manually for batch processing or set up automatic incremental processing
- The console provides guided setup with pre-built templates and ready-to-use rules
- For API access, comprehensive SDKs are available for Java, Python, JavaScript, and other languages
- AWS provides sample data, tutorials, and interactive scenarios to help users understand service capabilities before processing production data

Getting Started Guide:
https://docs.aws.amazon.com/entityresolution/latest/userguide/what-is-service.html

Data Removal

AWS Entity Resolution follows AWS standard data handling practices for offboarding. Customers maintain full control over their data as the service processes data from customer-controlled S3 buckets and writes results back to customer-specified locations. To remove data, customers delete their S3 input and output buckets, delete Entity Resolution workflows and schema mappings, and optionally delete associated AWS Glue Data Catalog entries. The service does not retain customer data beyond the processing period, and all intermediate processing data is automatically cleaned up after job completion. Customers can use AWS CloudTrail to audit all data access and deletion activities.

Backup/Restore and Disaster Recovery

AWS Entity Resolution does not provide built-in backup and disaster recovery capabilities as it is a processing service that operates on customer data stored in S3. Data durability and backup responsibilities lie with the underlying storage services customers use. AWS Entity Resolution workflows and configurations can be recreated using Infrastructure as Code tools like AWS CloudFormation or programmatically via APIs. Customers should implement their own backup strategies for S3 data and maintain copies of workflow configurations for disaster recovery purposes.

Backup Capabilities

AWS Entity Resolution does not have native backup capabilities and does not integrate directly with AWS Backup. The service processes data from customer S3 buckets and writes results back to customer-specified S3 locations:

- Customers are responsible for implementing backup strategies for their input data, output results, and workflow configurations
- Backup protection should be implemented at the S3 bucket level using S3 versioning, S3 replication, or AWS Backup policies applied to S3 resources

Disaster Recovery

AWS Entity Resolution does not directly integrate with AWS Elastic Disaster Recovery as it is a serverless processing service without persistent infrastructure. The service automatically runs across multiple Availability Zones within a region for high availability:

- For multi-region disaster recovery, customers must implement cross-region replication of their S3 data and recreate Entity Resolution workflows in secondary regions using Infrastructure as Code or API automation
- Workflow configurations and schema mappings can be exported and replicated across regions programmatically

Recovery Objectives

AWS Entity Resolution supports customer RPO and RTO objectives through its serverless, highly available architecture and flexible processing options. For RPO, customers can achieve low recovery point objectives by implementing cross-region S3 replication and maintaining current workflow configurations in multiple regions. For RTO, the service's serverless nature enables quick recovery as workflows can be rapidly recreated and executed without infrastructure provisioning delays. Near real-time matching capabilities support sub-second RTO requirements for critical matching operations.

Service Constraints

Service Quotas:
https://docs.aws.amazon.com/entityresolution/latest/userguide/quotas.html

FAQ: https://aws.amazon.com/entity-resolution/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/entityresolution/latest/userguide/

User Guide: https://docs.aws.amazon.com/entityresolution/latest/userguide/what-is-service.html

API Reference: https://docs.aws.amazon.com/entityresolution/latest/apireference/

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/entity-resolution/pricing/

Cost Optimization

AWS Entity Resolution offers several cost optimization strategies due to its pay-per-use pricing model based on records processed. Customers can optimize costs by implementing data preprocessing to reduce record volumes, using rule-based matching instead of ML-based matching when appropriate ($0.50 vs $0.10 per 1,000 records), and leveraging incremental processing to avoid reprocessing unchanged data:

- Additional optimizations include using data service provider-based matching for specific use cases, optimizing S3 storage classes for input and output data, implementing lifecycle policies for temporary processing data, and right-sizing processing jobs to avoid unnecessary resource consumption
- Regular monitoring of processing volumes and match rates helps identify opportunities for rule optimization and cost reduction

Savings Considerations

Primary cost savings come from the service's serverless architecture eliminating infrastructure management overhead and the pay-per-use model avoiding upfront costs. Customers save significantly compared to building custom entity resolution solutions by leveraging pre-built ML models, avoiding machine learning expertise requirements, and eliminating development and maintenance costs:

- Operational savings include reduced data engineering effort through automated schema mapping, decreased time-to-value with ready-to-use matching rules, and improved accuracy reducing downstream data cleanup costs
- Scale-based savings occur through automatic optimization of processing resources and the ability to process large datasets efficiently without manual tuning
- Organizations also benefit from reduced compliance and security implementation costs through built-in AWS security features and audit capabilities

Service Name

AWS Fargate

Service Overview

AWS Fargate is a serverless compute engine for containers that enables organizations to run containerized applications without managing servers, clusters, or underlying infrastructure. Compatible with both Amazon Elastic Container Service (ECS) and Amazon Elastic Kubernetes Service (EKS), Fargate allows developers to focus on application development while AWS handles infrastructure provisioning, scaling, and security. Each task or pod runs in its own isolation boundary with dedicated resources, providing enhanced security. Fargate supports both Linux and Windows containers with platform versions for Amazon Linux 2, Bottlerocket, and Microsoft Windows 2019 Server editions.

Key Use Cases

- Microservices Architecture: Deploy and scale microservices without infrastructure management overhead
- Batch Processing Workloads: Run batch jobs and data processing tasks with automatic resource provisioning
- Development and Test Environments: Quickly spin up containerized environments for development and testing
- Web Applications and APIs: Host containerized web applications with seamless scaling based on demand
- Stateful Applications: Run long-running stateful workloads without managing underlying infrastructure

Features

- **Serverless Container Compute**: Run containers without managing servers or clusters, with automatic infrastructure provisioning
- **Task Isolation**: Each task has its own isolation boundary with dedicated CPU, memory, and network resources
- **Automatic Scaling**: Scale applications based on demand without manual intervention or capacity planning
- **Fargate Spot**: Run interrupt-tolerant workloads at up to 70% discount using spare AWS compute capacity
- **Platform Version Support**: Support for Amazon Linux 2, Bottlerocket, and Windows 2019 Server platforms
- **Load Balancer Integration**: Seamless integration with Application Load Balancer and Network Load Balancer

- **Per-Second Billing**: Pay only for resources consumed with per-second billing granularity

Value Proposition

AWS Fargate eliminates the operational overhead of managing container infrastructure, allowing teams to focus on application development rather than server management. Key advantages include reduced total cost of ownership through pay-per-use pricing, enhanced security through task-level isolation, and simplified operations with automatic scaling:

- Fargate provides faster time-to-market by removing infrastructure provisioning delays and offers better resource utilization compared to traditional container deployments
- The service includes built-in high availability, compliance certifications for regulated industries, and seamless integration with the AWS ecosystem, making it ideal for organizations seeking serverless container solutions

Service Integration

AWS Fargate integrates deeply with Amazon ECS and Amazon EKS as the primary orchestration platforms. Core integrations include Amazon VPC for networking, IAM for security and access control, Amazon CloudWatch for monitoring and logging, AWS X-Ray for distributed tracing, and Amazon ECR for container image storage:

- Load balancing is provided through Application Load Balancer and Network Load Balancer
- Storage integrations include Amazon EFS for persistent storage, though Amazon EBS volumes cannot be mounted to Fargate tasks
- Additional integrations encompass AWS Lambda for event-driven architectures, Amazon SNS and SQS for messaging, and AWS Cloud Map for service discovery

Onboarding and Offboarding

Getting started with AWS Fargate requires creating an AWS account and setting up an IAM execution role for task execution. Customers can begin by creating a Fargate-enabled ECS cluster or EKS cluster through the AWS Management Console, AWS CLI, or infrastructure as code tools:

- The process involves defining task definitions with CPU and memory requirements, configuring networking settings including VPC and security groups, and deploying containers using pre-built images from Amazon ECR or other registries
- AWS provides step-by-step tutorials and getting started guides for both ECS and EKS implementations, along with sample applications and CloudFormation templates to accelerate deployment

Getting Started Guide:
https://docs.aws.amazon.com/AmazonECS/latest/developerguide/getting-started-fargate.html

Data Removal

Offboarding from AWS Fargate involves stopping all running tasks and services, deleting task definitions, and removing associated resources like load balancers and security groups. Data stored in Fargate containers is ephemeral and automatically deleted when tasks stop. For persistent data, customers must delete associated Amazon EFS file systems, DynamoDB tables, or other storage services. The process includes cleaning up IAM roles, removing CloudWatch log groups, and deleting ECS clusters or EKS clusters used with Fargate.

Backup/Restore and Disaster Recovery

AWS Fargate does not provide built-in backup capabilities as it is a serverless compute service for stateless containers. Data persistence and backup strategies depend on external storage services like Amazon EFS, Amazon S3, or Amazon RDS. Disaster recovery is achieved through multi-region deployments, container image replication across regions, and infrastructure as code practices. Organizations must implement application-level backup strategies and use AWS services like AWS Backup for supported resources.

Backup Capabilities

AWS Fargate itself does not have native backup capabilities since containers are ephemeral. Data backup must be implemented at the application level using AWS storage services. While Fargate tasks can write to Amazon EFS file systems which support AWS Backup integration, the Fargate service does not directly integrate with AWS Backup for task-level backups.

Disaster Recovery

AWS Fargate supports disaster recovery through multi-region deployment patterns but does not directly integrate with AWS Elastic Disaster Recovery. Organizations implement DR by deploying Fargate tasks across multiple Availability Zones and regions, using container image replication, and maintaining infrastructure as code templates for rapid environment recreation during disaster scenarios.

Recovery Objectives

AWS Fargate helps achieve low RTO objectives through rapid task startup capabilities, typically launching new tasks within minutes. RPO objectives depend on application-level data replication strategies to external storage services. The service supports automatic failover through Application Load Balancer health checks and can quickly spin up replacement tasks in different Availability Zones, contributing to overall disaster recovery capabilities.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/AmazonECS/latest/developerguide/service-quotas-manage.html

FAQ: https://aws.amazon.com/fargate/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/AmazonECS/latest/developerguide/AWS_Fargate.html

User Guide: https://docs.aws.amazon.com/AmazonECS/latest/developerguide/getting-started-fargate.html

API Reference: https://docs.aws.amazon.com/AmazonECS/latest/APIReference/

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/fargate/pricing/

Cost Optimization

AWS Fargate offers several unique cost optimization features including Fargate Spot for up to 70% savings on interrupt-tolerant workloads, per-second billing with no minimum commitments, and Compute Savings Plans for consistent usage patterns. Right-sizing recommendations through AWS Compute Optimizer help identify overprovisioned tasks:

- Graviton2-powered Fargate instances provide better price-performance ratios
- The service eliminates infrastructure management costs and reduces operational overhead through serverless architecture, allowing organizations to optimize spending by paying only for actual resource consumption

Savings Considerations

Primary cost savings come from eliminating server management overhead, reducing operational staff requirements, and paying only for consumed resources rather than provisioned capacity. Fargate Spot provides significant discounts for fault-tolerant workloads, while automatic scaling prevents overprovisioning:

- Organizations save on licensing costs by using Linux containers instead of Windows, and Graviton-based instances offer better price-performance

- The serverless model eliminates idle resource costs common in traditional container deployments, and integration with AWS Compute Optimizer enables continuous cost optimization through rightsizing recommendations

Service Name

AWS Fault Injection Service

Service Overview

AWS Fault Injection Service (AWS FIS) is a fully managed service that enables customers to perform fault injection experiments on AWS workloads based on chaos engineering principles. The service creates disruptive events to test application resilience by intentionally injecting faults such as stopping instances, disrupting network connectivity, stressing CPU/memory, or causing API errors. AWS FIS provides templates, controls, and guardrails needed to run experiments safely in production environments, including automatic rollback capabilities and stop conditions based on CloudWatch alarms. The service helps uncover application issues that are difficult to find through traditional testing methods.

Key Use Cases

- Use case 1: Application resilience testing by injecting faults like instance failures, network disruptions, and API errors to validate recovery procedures and failover mechanisms
- Use case 2: Chaos engineering implementation to build confidence in system reliability by regularly testing applications under real-world failure conditions in production environments
- Use case 3: Disaster recovery validation by simulating regional failures, availability zone disruptions, and service outages to verify backup systems and recovery time objectives

Features

- **Preconfigured Actions**: Ready-to-use fault injection activities for EC2, ECS, EKS, RDS, Lambda, S3, and other AWS services including instance termination, network disruption, and API errors
- **Stop Conditions**: CloudWatch alarm-based safety mechanisms that automatically halt experiments when predefined thresholds are exceeded to prevent uncontrolled damage
- **Multi-Account Experiments**: Capability to run fault injection experiments across multiple AWS accounts from a centralized orchestrator account with cross-account IAM roles
- **Systems Manager Integration**: Integration with SSM to execute custom fault injection scripts and automation documents for complex multi-step experiments
- **Experiment Templates**: Reusable blueprints that define actions, targets, stop conditions, and experiment configurations for consistent testing scenarios

Value Proposition

AWS FIS provides a fully managed chaos engineering solution that eliminates the complexity of building custom fault injection tools. Unlike traditional testing approaches, FIS offers production-safe experimentation with built-in guardrails, automatic rollback capabilities, and fine-grained targeting using AWS resource tags:

- The service integrates seamlessly with existing AWS services and provides preconfigured fault scenarios, reducing setup time and expertise requirements
- Key differentiators include comprehensive safety controls, multi-account support for enterprise scenarios, deep AWS service integration, and pay-per-use pricing that scales with experiment duration rather than requiring upfront infrastructure investment

Service Integration

AWS FIS integrates deeply with core AWS services to provide comprehensive fault injection capabilities. Primary integrations include Amazon CloudWatch for monitoring and stop conditions, AWS IAM for access control and experiment permissions, and AWS Systems Manager for custom script execution:

- The service works with compute services (EC2, ECS, EKS, Lambda), storage services (EBS, S3), databases (RDS, DynamoDB), and networking components (VPC, Route Tables, Transit Gateway)
- Additional integrations include AWS Application Recovery Controller for zonal autoshift testing, AWS Direct Connect for BGP session disruption, and multi-account management through AWS Organizations for enterprise-scale chaos engineering programs

Onboarding and Offboarding

Customers begin by creating IAM roles with necessary permissions for FIS to perform actions on target resources. The getting started process involves identifying target workloads, defining steady-state behavior, and creating experiment templates through the AWS console, CLI, or APIs:

- AWS provides pre-built experiment templates and tutorials for common scenarios like EC2 instance termination and network connectivity disruption
- Best practices include starting with non-production environments, implementing proper monitoring with CloudWatch alarms as stop conditions, and gradually expanding experiment scope
- The service requires no infrastructure provisioning as it's fully managed, allowing immediate experimentation once proper permissions are configured

Getting Started Guide: https://docs.aws.amazon.com/fis/latest/userguide/getting-started-planning.html

Data Removal

AWS FIS automatically retains completed experiment data for 120 days in each region, after which it's automatically deleted. During offboarding, customers should delete experiment templates and ensure no active experiments are running. The service doesn't store persistent customer data beyond experiment logs and templates, so offboarding primarily involves removing IAM roles and cleaning up experiment artifacts. Customers can manually delete experiment templates and associated CloudWatch alarms if no longer needed.

Backup/Restore and Disaster Recovery

AWS FIS itself doesn't provide backup and disaster recovery capabilities as it's a testing service that operates on existing resources. However, the service is instrumental in validating backup and disaster recovery procedures by simulating failures and testing recovery mechanisms. FIS experiments help verify that backup systems, failover processes, and disaster recovery plans function correctly under stress conditions, ensuring RPO and RTO objectives are met.

Backup Capabilities

AWS FIS doesn't have traditional backup capabilities as it's a fault injection testing service rather than a data storage service. The service doesn't integrate directly with AWS Backup since it operates on existing resources to test their resilience rather than storing data that requires backup protection.

Disaster Recovery

AWS FIS doesn't support AWS Elastic Disaster Recovery integration for its own operations, but serves as a critical tool for testing disaster recovery scenarios. The service helps validate disaster recovery procedures by simulating regional failures, AZ disruptions, and service outages to ensure recovery mechanisms function properly.

Recovery Objectives

AWS FIS helps customers validate their RPO and RTO objectives by simulating real failure scenarios and measuring actual recovery times. Through controlled chaos experiments, customers can verify that backup restoration meets RPO requirements and failover procedures achieve target RTO windows. The service enables testing of automated recovery mechanisms and identification of gaps between planned and actual recovery performance.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/fis/latest/userguide/fis-quotas.html

FAQ: https://aws.amazon.com/fis/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/fis/latest/userguide/what-is.html

User Guide: https://docs.aws.amazon.com/fis/latest/userguide/

API Reference: https://docs.aws.amazon.com/fis/latest/APIReference/

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/fis/pricing/

Cost Optimization

AWS FIS offers unique cost optimization through its pay-per-use pricing model charged per action-minute rather than provisioned infrastructure. Customers pay $0.10 per action-minute in most regions, with additional charges only for multi-account experiments:

- Cost optimization strategies include running shorter experiments, targeting specific resources rather than broad sweeps, and using experiment templates to avoid repetitive setup costs
- The service eliminates infrastructure costs for chaos engineering tools while providing comprehensive testing capabilities
- Experiment reports cost $5 each, allowing customers to balance detailed reporting needs with cost considerations
- Regional pricing variations exist, with GovCloud regions priced at $0.12 per action-minute

Savings Considerations

Primary cost savings come from preventing costly production outages through proactive resilience testing, reducing mean time to recovery (MTTR) by validating automation and procedures. By identifying weaknesses before they cause customer-impacting incidents, organizations avoid revenue loss, SLA penalties, and emergency response costs:

- FIS eliminates the need to build and maintain custom chaos engineering infrastructure, reducing operational overhead and tooling costs
- The service's precise targeting capabilities minimize unnecessary resource consumption during testing
- Multi-account support reduces the need for multiple testing environments across different teams, consolidating chaos engineering activities and associated costs while improving organizational learning and best practice sharing

Service Name

AWS Firewall Manager

Service Overview

AWS Firewall Manager is a security management service that simplifies administration and maintenance tasks across multiple accounts and resources for various protections, including AWS WAF, AWS Shield Advanced, Amazon VPC security groups and network ACLs, AWS Network Firewall, and Amazon Route 53 Resolver DNS Firewall. The service enables centralized configuration and management of firewall rules and protections across accounts and applications within an AWS Organizations structure. Users set up protections once, and Firewall Manager automatically applies them across all accounts and resources, including newly added ones, ensuring consistent security policy enforcement organization-wide.

Key Use Cases

- Centralized security policy management: Deploy consistent firewall rules and protections across multiple AWS accounts in an organization from a single management point
- Compliance enforcement: Automatically monitor for new resources and ensure they comply with mandatory security policies from day one of deployment
- Multi-layer protection orchestration: Coordinate AWS WAF, Shield Advanced, Network Firewall, VPC security groups, and DNS Firewall rules across organizational boundaries
- Security audit and remediation: Audit existing security configurations for over-permissive rules and automatically remediate non-compliant resources

Features

- **Multi-service protection management**: Centrally manages AWS WAF, AWS Shield Advanced, VPC security groups, Network ACLs, AWS Network Firewall, and Route 53 Resolver DNS Firewall protections
- **Cross-account policy enforcement**: Applies security policies automatically across multiple AWS accounts within an AWS Organizations structure
- **Automatic resource protection**: Automatically applies protections to new resources as they are created, ensuring continuous compliance
- **Hierarchical rule management**: Supports hierarchical rule enforcement allowing common organizational rules while permitting account-specific customizations
- **Compliance dashboard**: Provides centralized visibility into compliance status with notifications for non-compliant resources and threat monitoring

- **Third-party firewall integration**: Supports integration with third-party firewall solutions from AWS Marketplace including Palo Alto Networks and Fortigate

Value Proposition

AWS Firewall Manager provides unparalleled centralized security management for organizations with multiple AWS accounts, eliminating the complexity of manually configuring security policies across accounts. The service automatically ensures consistent security posture as organizations scale, reducing administrative overhead and human error:

- Key advantages include automatic policy application to new resources, hierarchical rule management that balances central control with local flexibility, and comprehensive compliance monitoring
- Organizations benefit from reduced security management complexity, improved security consistency, and the ability to leverage managed rules from AWS Marketplace while maintaining centralized oversight of their entire AWS security infrastructure

Service Integration

AWS Firewall Manager deeply integrates with AWS Organizations as a prerequisite, requiring organizational structure for multi-account management. Core service integrations include AWS WAF for web application protection, AWS Shield Advanced for DDoS protection, AWS Network Firewall for network-level filtering, Amazon VPC for security groups and network ACLs, and Amazon Route 53 Resolver for DNS filtering:

- The service requires AWS Config for continuous resource monitoring and compliance tracking, and integrates with AWS Security Hub for centralized security findings management
- Additional integrations include AWS Resource Access Manager for cross-account resource sharing, AWS CloudTrail for API logging, and AWS Marketplace for third-party security solutions

Onboarding and Offboarding

Getting started with AWS Firewall Manager requires completing several prerequisite steps in sequence. First, organizations must join AWS Organizations and enable all features:

- Next, they must designate a Firewall Manager administrator account from either the management account or a delegated member account
- AWS Config must be enabled across all member accounts for continuous resource monitoring
- Optionally, customers can enable AWS Resource Access Manager for Network Firewall and DNS Firewall resource sharing

- Once prerequisites are complete, customers create their first security policy by selecting the protection type (WAF, Shield, Network Firewall, etc.), configuring rules and protections, defining policy scope including accounts and resources, and choosing between automatic or manual remediation modes

Getting Started Guide: https://docs.aws.amazon.com/waf/latest/developerguide/fms-prereq.html

Data Removal

Offboarding from AWS Firewall Manager involves deleting policies using the DeletePolicy API with the DeleteAllPolicyResources parameter set to true for complete cleanup. This removes rule groups, disassociates web ACLs from resources, deletes Firewall Manager-created security groups, and cleans up Network Firewall resources. For third-party firewall integrations, the DisassociateThirdPartyFirewall API removes all associated firewalls and tenant configurations. Customer data is removed when policies are deleted, though underlying service resources may require separate cleanup actions depending on the protection type.

Backup/Restore and Disaster Recovery

AWS Firewall Manager does not provide direct backup and disaster recovery capabilities as it is a management service for security policies rather than a data storage service. The service maintains policy configurations and compliance state information, but does not offer specific backup features. Disaster recovery relies on the underlying AWS infrastructure resilience, and policies can be recreated using infrastructure as code practices or by reconfiguring through the console or APIs.

Backup Capabilities

AWS Firewall Manager does not have built-in backup capabilities and does not integrate directly with AWS Backup service. As a security policy management service, it focuses on configuration management rather than data protection. Policy configurations should be backed up through infrastructure as code practices, version control systems, or by maintaining documented configurations that can be recreated.

Disaster Recovery

AWS Firewall Manager does not directly integrate with AWS Elastic Disaster Recovery as it is a management service rather than a compute or data service. Disaster recovery for Firewall Manager relies on AWS infrastructure resilience and the ability to recreate policies in alternate regions. The service operates regionally, so organizations must create separate policies in each region for comprehensive coverage.

Recovery Objectives

AWS Firewall Manager supports recovery objectives by enabling rapid policy recreation and deployment across multiple accounts and regions. While not directly providing RPO/RTO capabilities, the service can quickly restore security postures by reapplying policies to resources. Organizations can achieve low RTO by maintaining policy configurations in infrastructure as code and leveraging the service's automatic policy application capabilities for rapid security restoration.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/waf/latest/developerguide/fms-limits.html

FAQ: https://aws.amazon.com/firewall-manager/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/waf/latest/developerguide/fms-chapter.html

User Guide: https://docs.aws.amazon.com/waf/latest/developerguide/fms-chapter.html

API Reference: https://docs.aws.amazon.com/fms/2018-01-01/APIReference/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/firewall-manager/pricing/

Cost Optimization

AWS Firewall Manager pricing is based on policy types and regional deployment, with charges primarily for underlying services rather than the management layer itself. Key cost optimization strategies include leveraging the fact that AWS Shield Advanced customers receive Firewall Manager WAF policies at no additional charge:

- Organizations can optimize costs by carefully scoping policies to avoid unnecessary resource inclusion and utilizing manual remediation mode during testing phases
- Consolidating similar protection requirements into fewer policies and leveraging managed rules from AWS Marketplace can provide cost-effective security coverage
- Understanding that charges are primarily from underlying services allows for targeted optimization of WAF, Network Firewall, and Config usage

Savings Considerations

Primary cost savings come from avoiding duplicate security configurations across multiple accounts and reducing administrative overhead through centralized management. Organizations save on operational costs by eliminating manual security policy deployment and maintenance across accounts:

- The service enables bulk purchasing efficiencies for Shield Advanced subscriptions across organizational accounts
- Automated compliance monitoring reduces the need for manual security auditing resources
- By preventing security misconfigurations through consistent policy application, organizations avoid potential costs associated with security incidents
- Leveraging AWS Marketplace managed rules through Firewall Manager can provide economies of scale compared to developing custom security rules for each account separately

Service Name

AWS Global Accelerator

Service Overview

AWS Global Accelerator is a networking service that improves the availability and performance of applications for both local and global users by routing traffic over the AWS global network. It provides static IP addresses that serve as fixed entry points to applications and automatically routes user traffic to optimal endpoints based on performance, health, and policies. The service utilizes a global network of 130 Points of Presence (PoPs) in 95 cities across 53 countries to accelerate application delivery and reduce latency by up to 49% and jitter by up to 58%.

Key Use Cases

- Latency-sensitive applications: Gaming, media, mobile apps, ad-tech, and financial applications that require reduced internet latency and jitter for optimal user experience
- Multi-Region disaster recovery: Detecting application endpoint failures in the primary AWS Region and instantly rerouting traffic to the next available, closest AWS Region for business continuity
- Global application scaling: Industries like IoT, retail, media, automotive, and healthcare where frequent client application updates are not feasible but global reach is required
- Custom routing for specialized applications: VoIP, online gaming, and EdTech applications that require custom logic to map users to specific endpoints while maintaining performance benefits

Features

- **Static Anycast IP Addresses**: Provides two static IPv4 addresses and two static IPv6 addresses (for dual-stack) that serve as fixed entry points to applications, with support for Bring Your Own IP (BYOIP)
- **Global Performance-Based Routing**: Routes TCP and UDP traffic to healthy application endpoints in the closest AWS Region based on user location, endpoint health, and policies
- **Health Monitoring**: Continuously monitors endpoint health using TCP, HTTP, and HTTPS health checks and automatically redirects traffic to healthy endpoints
- **Traffic Control**: Fine-grained traffic control with traffic dials to adjust percentage of traffic to specific AWS Regions and endpoint weights for load distribution

- **DDoS Protection**: Protected by AWS Shield Standard with option to enable AWS Shield Advanced for enhanced DDoS protection and 24x7 access to DDoS Response Team
- **Custom Routing Accelerators**: Allows custom application logic to route user traffic to specific Amazon EC2 instance destinations in single or multiple AWS Regions
- **Client Affinity**: Enables building applications that require maintaining state by consistently routing users to the same endpoint
- **Fault-Tolerant Design**: Two static IP addresses serviced by independent network zones with automatic failover capabilities

Value Proposition

AWS Global Accelerator provides unique advantages over alternatives by combining global network acceleration with simplified management. Unlike DNS-based solutions, it provides static IP addresses that eliminate the complexity of managing region-specific IPs and DNS updates:

- Compared to Amazon CloudFront, Global Accelerator works with non-HTTP protocols and provides deterministic fast regional failover
- It offers up to 60% performance improvement through the AWS global network, reduces first byte latency by up to 49%, and provides instant failover capabilities
- The service integrates seamlessly with existing AWS infrastructure including ELB, EC2, and VPC resources while providing built-in DDoS protection and requiring no long-term commitments

Service Integration

AWS Global Accelerator integrates deeply with core AWS services to provide comprehensive application acceleration. It works with Elastic Load Balancing (Application Load Balancers and Network Load Balancers) for regional load distribution, Amazon EC2 instances and Elastic IP addresses as direct endpoints, and Amazon VPC subnets for custom routing scenarios:

- The service integrates with Amazon Route 53 for DNS management and custom domain routing, AWS Shield for DDoS protection, Amazon CloudWatch for monitoring and logging, and AWS IAM for access control and cross-account resource sharing
- Additional integrations include Amazon S3 Multi-Region Access Points for object storage acceleration and AWS Application Recovery Controller for coordinated failover management

Onboarding and Offboarding

Getting started with AWS Global Accelerator involves three simple steps through the AWS Management Console or API. First, create an accelerator by providing a name, choosing between standard or custom routing types, and selecting IPv4 or dual-stack address types:

- Second, configure listeners to process inbound connections based on protocol and port specifications, then set up endpoint groups associated with specific AWS Regions
- Third, register endpoints such as Application Load Balancers, Network Load Balancers, EC2 instances, or Elastic IP addresses, configuring weights and health check settings
- The service must be managed from the US West (Oregon) Region, and users can optionally bring their own IP addresses (BYOIP) for custom requirements

Getting Started Guide: https://docs.aws.amazon.com/global-accelerator/latest/dg/getting-started.html

Data Removal

To offboard from AWS Global Accelerator, customers must first disable the accelerator through the console or API, then remove all associated listeners and endpoint groups before deletion. When an accelerator is deleted, the static IP addresses assigned to it are permanently lost and cannot be recovered. Cross-account attachments can be deleted separately, which revokes permissions for shared resources. The deletion process is irreversible, so customers should ensure proper planning before removing accelerators to avoid losing static IP configurations.

Backup/Restore and Disaster Recovery

AWS Global Accelerator itself is a disaster recovery enablement service rather than providing traditional backup capabilities. It supports disaster recovery by instantly detecting application endpoint failures and automatically rerouting traffic to healthy endpoints in alternate AWS Regions. The service enables multi-Region resiliency with automatic failover capabilities, helping customers achieve rapid recovery objectives. Configuration settings and accelerator definitions can be managed through Infrastructure as Code tools like AWS CloudFormation for consistent deployment and recovery of the service configuration itself.

Backup Capabilities

AWS Global Accelerator does not integrate with AWS Backup as it is a network routing service rather than a data storage service. The service configuration can be backed up through Infrastructure as Code practices using AWS CloudFormation templates or API scripts:

- Static IP addresses are maintained by the service as long as the accelerator exists but are lost upon deletion
- Customers should maintain configuration backups through code repositories and automation scripts to quickly recreate accelerators if needed

Disaster Recovery

AWS Global Accelerator enhances disaster recovery capabilities but does not directly integrate with AWS Elastic Disaster Recovery. Instead, it serves as a traffic management layer that works in conjunction with disaster recovery solutions by providing instant traffic rerouting to healthy endpoints:

- The service supports automated failover at the network level, complementing application-level disaster recovery solutions
- It can be used with AWS Application Recovery Controller for coordinated failover management across multiple Regions and services

Recovery Objectives

AWS Global Accelerator significantly improves Recovery Time Objectives (RTO) by providing instant traffic failover to healthy endpoints within seconds of detecting failures. The service continuously monitors endpoint health and immediately redirects traffic when issues are detected, enabling near-zero RTO for network-level failovers. For Recovery Point Objectives (RPO), Global Accelerator works with underlying services and applications to minimize data loss during failover scenarios. The service's anycast IP addresses and global routing capabilities ensure users are always directed to the nearest available healthy endpoint.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/global-accelerator/latest/dg/limits-global-accelerator.html

FAQ: https://aws.amazon.com/global-accelerator/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/global-accelerator/latest/dg/what-is-global-accelerator.html

User Guide: https://docs.aws.amazon.com/global-accelerator/latest/dg/what-is-global-accelerator.html

API Reference: https://docs.aws.amazon.com/global-accelerator/latest/api/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/global-accelerator/pricing/

Cost Optimization

AWS Global Accelerator offers several cost optimization strategies including traffic dials to control the percentage of traffic directed to specific AWS Regions, enabling gradual traffic shifting and cost management. Customers can optimize costs by strategically placing endpoints in regions with lower data transfer costs and using traffic weights to distribute load efficiently:

- The service supports Bring Your Own IP (BYOIP) to avoid additional IP address costs
- Traffic routing optimization can reduce overall data transfer costs by directing users to the nearest endpoints, and the pay-as-you-go model with no long-term commitments allows for flexible cost management based on actual usage patterns

Savings Considerations

Cost savings with AWS Global Accelerator come from improved application performance leading to better user experience and potentially reduced infrastructure needs in multiple regions. The service eliminates the need for complex DNS-based traffic management solutions and reduces operational overhead:

- By using traffic dials for blue/green deployments and A/B testing, customers can optimize resource utilization and avoid over-provisioning
- The fixed hourly fee model ($0.025 per hour) combined with usage-based Data Transfer-Premium charges allows predictable cost planning
- Performance improvements can lead to reduced compute requirements and better resource utilization across endpoints, potentially offsetting the service costs through overall infrastructure efficiency gains

Service Name

AWS Glue

Service Overview

AWS Glue is a serverless data integration service that simplifies data discovery, preparation, movement, and integration for analytics, machine learning, and application development. It supports over 70 data sources and offers a centralized data catalog with tools for authoring, running, and monitoring jobs. AWS Glue consolidates data integration capabilities including data discovery, modern ETL, cleansing, transforming, and centralized cataloging. It integrates with AWS analytics services and Amazon S3 data lakes, providing both visual and code-based interfaces through AWS Glue Studio. The service automatically scales based on workload, handles infrastructure provisioning, and offers pay-as-you-go billing with generative AI assistance for ETL authoring and troubleshooting.

Key Use Cases

- Data lake creation and management: Building modern data platforms with unified metadata repository and automated data discovery
- ETL pipeline development: Building complex ETL pipelines for batch, micro-batch, and streaming data processing with visual or code-based interfaces
- Data warehousing and analytics: Preparing and loading data from various sources into data warehouses like Amazon Redshift for business intelligence
- Machine learning data preparation: Cleaning, transforming, and preparing data for ML workflows with built-in data quality features
- Real-time data processing: Processing streaming data from Amazon Kinesis, Apache Kafka, and Amazon MSK with in-flight transformations
- Data catalog and governance: Creating centralized metadata repository with schema management, data quality monitoring, and access controls

Features

- **AWS Glue Data Catalog**: Persistent metadata store serving as central repository for data assets with automatic schema discovery and version history
- **AWS Glue Studio**: Visual interface for creating, running, and monitoring data integration jobs with drag-and-drop functionality
- **Crawlers and Classifiers**: Automatically scan data repositories, classify data, extract schema information, and populate metadata catalog
- **ETL Jobs System**: Managed infrastructure for running Scala or PySpark scripts with automatic scaling and event-driven triggers
- **Streaming ETL**: Real-time data processing using Apache Spark Structured Streaming for continuous data transformation

- **AWS Glue DataBrew**: No-code visual data preparation with over 250 ready-made transformations for data cleaning and normalization
- **Data Quality Management**: Built-in data quality rules, anomaly detection, and automated data cleansing with machine learning
- **Schema Registry**: Validation and control of streaming data using registered Apache Avro schemas with version management
- **Interactive Sessions**: Jupyter notebook environment for editing, debugging, and testing ETL code with fast startup times
- **Generative AI Integration**: AI-powered assistance for ETL code generation, Apache Spark job modernization, and troubleshooting

Value Proposition

AWS Glue provides serverless data integration that eliminates infrastructure management overhead while offering comprehensive data processing capabilities. Customers benefit from automatic scaling, pay-per-use pricing, and reduced operational complexity compared to traditional ETL solutions:

- The service accelerates data pipeline development through visual interfaces and generative AI assistance, reducing development time by up to 80% for data preparation tasks
- AWS Glue's tight integration with the AWS ecosystem enables seamless connectivity to over 100 data sources and native compatibility with analytics services like Amazon Athena, Redshift, and QuickSight
- The centralized Data Catalog provides unified metadata management across all data assets, enabling better data governance and discovery

Service Integration

AWS Glue deeply integrates with core AWS analytics and storage services including Amazon S3 for data lakes, Amazon Redshift for data warehousing, and Amazon Athena for query services. The Data Catalog serves as a drop-in replacement for Apache Hive Metastore and integrates with Amazon EMR, Amazon QuickSight, and AWS Lake Formation:

- Streaming capabilities connect with Amazon Kinesis Data Streams, Amazon MSK (Managed Streaming for Apache Kafka), and AWS Lambda for event-driven architectures
- Security integration includes AWS IAM for access control, AWS KMS for encryption, and VPC for network isolation
- The service also integrates with Amazon SageMaker for machine learning workflows, AWS CloudWatch for monitoring, and AWS CloudTrail for auditing

Onboarding and Offboarding

Customers can start using AWS Glue by setting up IAM permissions and accessing the AWS Glue console or AWS Glue Studio. The typical onboarding process involves creating a default database in the Data Catalog, configuring network access to data sources, and optionally setting up encryption:

- Users can begin with crawlers to automatically discover and catalog existing data, then create ETL jobs using either the visual interface in AWS Glue Studio or code-based approaches
- AWS provides getting started guides, tutorials, and sample datasets to help customers build their first data integration pipelines
- The service offers interactive sessions for iterative development and testing before moving to production workloads

Getting Started Guide: https://docs.aws.amazon.com/glue/latest/dg/setting-up.html

Data Removal

AWS Glue data removal involves deleting ETL jobs, crawlers, and associated metadata from the Data Catalog. Customers can remove table definitions, database objects, and connection configurations through the console or API. Job history and logs stored in CloudWatch require separate deletion. Data processed by Glue jobs remains in customer-controlled storage locations like S3 buckets and must be deleted independently. Complete offboarding requires removing IAM roles, security groups, and any VPC endpoints created specifically for Glue operations.

Backup/Restore and Disaster Recovery

AWS Glue provides built-in high availability across multiple Availability Zones within a region. The Data Catalog metadata is automatically backed up and replicated by AWS with version history tracking for schema changes. ETL job definitions and configurations are stored redundantly and can be exported for backup purposes. However, customers are responsible for backing up their actual data in connected data stores like S3 buckets or databases. Cross-region replication of metadata catalogs requires manual setup through APIs or infrastructure as code.

Backup Capabilities

AWS Glue automatically backs up the Data Catalog metadata with built-in redundancy and version control. Job definitions, crawler configurations, and connection details are persistently stored with high durability:

- The service does not directly integrate with AWS Backup service since it primarily manages metadata rather than customer data

- Customers can export job scripts and configurations for additional backup through APIs or CloudFormation templates

Disaster Recovery

AWS Glue supports disaster recovery through its serverless, multi-AZ architecture that automatically handles infrastructure failures. The service does not directly integrate with AWS Elastic Disaster Recovery as it manages metadata and job orchestration rather than persistent compute resources. Customers can implement cross-region disaster recovery by replicating job definitions and Data Catalog metadata to secondary regions using APIs or infrastructure automation tools.

Recovery Objectives

AWS Glue helps meet RPO objectives through automatic metadata backup and version control in the Data Catalog, minimizing potential data loss. For RTO objectives, the serverless architecture enables rapid job restart and automatic scaling without infrastructure provisioning delays. Cross-region job deployment can further reduce RTO by enabling failover to secondary regions. However, overall recovery times depend on the underlying data stores and network connectivity to data sources.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/general/latest/gr/glue.html

FAQ: https://aws.amazon.com/glue/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/glue/latest/dg/what-is-glue.html

User Guide: https://docs.aws.amazon.com/glue/latest/dg/

API Reference: https://docs.aws.amazon.com/glue/latest/webapi/

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/glue/pricing/

Cost Optimization

AWS Glue offers several cost optimization strategies including flexible job execution for non-urgent workloads, autoscaling capabilities that dynamically adjust DPU allocation based on workload demands, and pay-per-use pricing with no upfront costs. Customers can optimize costs by choosing appropriate worker types (G.1X, G.2X, G.025X) based on

job requirements, implementing job bookmarks to process only new data, and using spot instances for development workloads:

- Data partitioning and efficient file formats like Parquet can reduce processing time and costs
- The service also provides usage monitoring through CloudWatch to identify optimization opportunities and right-size resource allocation

Savings Considerations

Primary cost savings come from the serverless model eliminating infrastructure provisioning and management overhead compared to self-managed ETL solutions. Automatic scaling prevents over-provisioning while ensuring performance during peak workloads:

- The Data Catalog reduces operational costs by providing centralized metadata management and automatic schema discovery, eliminating manual cataloging efforts
- Built-in data quality features and machine learning capabilities reduce the need for additional tools and custom development
- Integration with other AWS services enables cost-effective end-to-end data pipelines without data transfer charges within the same region
- Customers typically see 20-40% cost reduction compared to traditional ETL platforms when factoring in operational overhead savings

Service Name

AWS Ground Station

Service Overview

AWS Ground Station is a fully managed service that enables customers to control satellite communications, process satellite data, and scale operations without building or managing ground station infrastructure. The service provides secure, fast, and predictable satellite communications across a global infrastructure using AWS-managed antennas and ground stations. It eliminates the need for customers to build, manage, or scale their own ground station infrastructure, allowing them to focus on developing innovative applications using satellite data. AWS Ground Station uses a low-latency, high-bandwidth global fiber network to process satellite data within seconds of reception.

Key Use Cases

- Real-time satellite data downlinking and storage: Downlink satellite data in real-time or store it directly in Amazon S3 for immediate processing and analysis
- Satellite command and control operations: Uplink commands and data to satellites while performing Telemetry, Tracking, and Command (TT&C) functions for mission operations
- Federal and defense space resilience: Support government and defense organizations with secure, scalable ground segment solutions for national security and space operations
- Earth observation and remote sensing: Process large amounts of satellite imagery and sensor data for weather monitoring, agriculture, and environmental analysis
- Software Defined Radio (SDR) applications: Utilize SDR and Front End Processor (FEP) capabilities for advanced signal processing and custom satellite communication protocols

Features

- **Fully Managed Global Ground Station Network**: AWS-managed antennas and ground stations integrated with AWS Global Infrastructure across multiple regions worldwide
- **Multiple Frequency Band Support**: Supports X-band, S-band, and Ku-band satellite protocols with various modulation and encoding schemes
- **Pay-per-Minute Pricing**: Purchase antenna time in one-minute increments with no upfront costs or long-term commitments
- **Wideband Digital Intermediate Frequency (DigIF)**: High-bandwidth signal processing capabilities for advanced satellite communications at select sites

- **Digital Twin Testing**: Test contact orchestration workflows against virtual ground stations for reduced costs and risk-free validation
- **Automatic Contact Scheduling**: Automated satellite communication session scheduling and management with up to 30 days advance booking
- **Customer-Provided Ephemerides**: Support for custom ephemeris data upload for antenna targeting during LEOP or orbital maneuvers
- **AWS Service Integration**: Direct integration with Amazon S3, EC2, Lambda, SageMaker, and other AWS services for data processing

Value Proposition

AWS Ground Station eliminates the complexity and cost of building traditional ground station infrastructure while providing up to 80% cost savings on ground station operations. Customers benefit from global reach without geographical constraints, pay-only-for-use pricing model, and seamless integration with the full AWS ecosystem for data processing and analytics:

- The service offers unparalleled flexibility with no long-term commitments, automatic scaling capabilities, and the ability to focus on mission objectives rather than infrastructure management
- Built-in security features, encryption, and compliance capabilities make it suitable for government and commercial applications

Service Integration

AWS Ground Station integrates natively with core AWS services to enable comprehensive satellite data workflows. Amazon S3 provides automatic data storage and archiving, while Amazon EC2 handles satellite command processing and data analysis:

- Amazon Kinesis Data Streams manages real-time data ingestion from satellites, and AWS Lambda enables serverless data processing and format conversions
- Amazon SageMaker supports machine learning applications on satellite data, while Amazon VPC ensures secure networking
- AWS CloudWatch provides monitoring and telemetry, AWS IAM manages access control, and Amazon SNS delivers notifications
- AWS KMS provides encryption services for data security

Onboarding and Offboarding

Customers begin by signing up for an AWS account and setting up appropriate IAM permissions for Ground Station access. The onboarding process involves satellite registration through data collection, technical validation, spectrum licensing, and integration testing:

- Customers must create mission profiles with tracking configurations and dataflows, then configure AWS resources like VPC, S3, and EC2 for data processing
- Contact scheduling requires valid ephemeris data and can be done up to 30 days in advance
- AWS provides comprehensive documentation, getting started guides, and console-based management tools to streamline the setup process

Getting Started Guide: https://docs.aws.amazon.com/ground-station/latest/ug/getting-started.html

Data Removal

The offboarding process involves canceling scheduled contacts, deleting mission profiles and satellite configurations, and removing dataflow endpoint groups. Customers can delete stored satellite data from Amazon S3 buckets and terminate associated EC2 instances. All Ground Station resources can be removed through the AWS console, CLI, or APIs, with no long-term commitments requiring formal contract termination.

Backup/Restore and Disaster Recovery

AWS Ground Station leverages AWS's inherent disaster recovery capabilities through its global infrastructure. Satellite data can be automatically replicated across regions using Amazon S3 cross-region replication. Mission profiles and configurations can be recreated in alternate regions using Infrastructure as Code. The service's distributed nature provides built-in redundancy through multiple ground station locations.

Backup Capabilities

AWS Ground Station does not have native backup capabilities but integrates with AWS Backup for protecting associated resources like EC2 instances and EBS volumes. Satellite data stored in Amazon S3 can leverage S3 versioning, cross-region replication, and S3 Glacier for long-term archival. Mission configurations should be backed up using Infrastructure as Code practices.

Disaster Recovery

AWS Ground Station supports disaster recovery through its multi-region architecture and integration with AWS services. The service does not directly integrate with AWS Elastic Disaster Recovery but benefits from the underlying AWS infrastructure's resilience. Customers can implement disaster recovery strategies using cross-region data replication and automated failover mechanisms.

Recovery Objectives

AWS Ground Station helps meet RPO and RTO objectives through its global infrastructure and real-time data processing capabilities. Cross-region data replication can achieve near-zero RPO for critical satellite data, while the distributed ground station network

provides multiple communication paths for improved availability. Automated contact scheduling and digital twin testing help ensure operational continuity.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/general/latest/gr/gs.html

FAQ: https://aws.amazon.com/ground-station/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/ground-station/latest/ug/index.html

User Guide: https://docs.aws.amazon.com/ground-station/latest/ug/index.html

API Reference: https://docs.aws.amazon.com/ground-station/latest/APIReference/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/ground-station/pricing/

Cost Optimization

AWS Ground Station offers unique cost optimization through pay-per-minute pricing with no upfront infrastructure investment, potentially saving up to 80% compared to traditional ground stations. Reserved scheduling provides discounted rates for regular users with monthly commitments over 12 months:

- The digital twin feature allows cost-effective testing and validation without using actual antenna time
- Customers can optimize costs by scheduling contacts during off-peak times, using narrow-band communications when possible, and leveraging AWS's global network to select the most cost-effective ground station locations

Savings Considerations

Primary cost savings come from eliminating capital expenditure on ground station infrastructure, maintenance, and staffing. The pay-as-you-go model aligns costs directly with usage, avoiding fixed infrastructure costs:

- Reserved pricing offers additional discounts for predictable workloads, while the global network eliminates the need for multiple geographic installations

- Integration with AWS services reduces data transport and processing costs through optimized data paths
- Automated scheduling and management reduce operational overhead, and the shared infrastructure model provides economies of scale unavailable to private ground stations

Service Name

AWS HealthImaging

Service Overview

AWS HealthImaging is a HIPAA-eligible cloud service designed for healthcare providers, life science organizations, and their software partners to store, analyze, and share medical images at petabyte scale. The service provides cloud-native medical imaging and picture archiving and communication system (PACS) capabilities, supporting DICOM P10 format data with APIs for low-latency retrieval and purpose-built storage. It enables healthcare organizations to reduce infrastructure costs by running medical imaging applications at scale from a single, authoritative copy of each medical image in AWS Cloud while maintaining performance previously only possible on-premises.

Key Use Cases

- Enterprise imaging: Store and stream medical imaging data directly from AWS Cloud while preserving low-latency performance and high availability
- Long-term image archival: Save costs on long-term image archival while maintaining subsecond image retrieval access
- AI/ML development: Run artificial intelligence and machine learning inference over imaging archives with support from other AWS services
- Multimodal analysis: Combine clinical imaging data with AWS HealthLake health data and AWS HealthOmics omics data for precision medicine insights

Features

- **Developer-friendly DICOM metadata**: Returns DICOM metadata in developer-friendly format with human-friendly keywords rather than hexadecimal numbers, with normalized Patient, Study, and Series level elements
- **SIMD-accelerated image decoding**: Returns image frames encoded as High Throughput JPEG 2000 (HTJ2K) with SIMD acceleration, delivering order-of-magnitude faster performance than JPEG2000
- **Pixel data verification**: Built-in pixel data verification using IFRC checksums to ensure accuracy of decoded images during import process
- **Scalable DICOM imports**: Cloud-native parallel import capabilities for multiple DICOM studies without impacting clinical workloads
- **Service-managed DICOM data hierarchy**: Automatically organizes DICOM P10 data by Patient, Study, and Series levels into image sets corresponding to DICOM series
- **DICOMweb API compatibility**: Offers DICOMweb conformant APIs for integration with existing applications plus cloud-native APIs for advanced operations

- **Intelligent tiering**: Automatic movement between Frequent Access and Archive Instant Access tiers based on usage patterns to optimize costs

Value Proposition

AWS HealthImaging offers industry-leading performance with HTJ2K encoding that is an order of magnitude faster than JPEG2000 and twice as fast as other DICOM transfer syntaxes. The service provides HIPAA eligibility, built-in security features, and seamless integration with the broader AWS ecosystem for advanced analytics and machine learning:

- Unlike traditional PACS solutions, HealthImaging eliminates infrastructure management overhead while providing petabyte-scale storage, automatic cost optimization through intelligent tiering, and developer-friendly APIs that accelerate application development
- The service ensures data integrity through pixel data verification and offers both DICOMweb compatibility for existing integrations and cloud-native APIs for modern applications

Service Integration

AWS HealthImaging integrates deeply with AWS Identity and Access Management for secure resource access, Amazon S3 for DICOM data staging and import, Amazon CloudWatch for monitoring and observability, AWS CloudTrail for API usage tracking, and Amazon EventBridge for event-driven architectures. The service leverages AWS PrivateLink for secure VPC connectivity without internet exposure, AWS CloudFormation for infrastructure as code deployments, and integrates with AWS HealthLake for health data and AWS HealthOmics for omics data to enable comprehensive multimodal analysis. Additional integrations include Amazon SageMaker for AI/ML capabilities and Amazon Cognito for authentication in viewer applications.

Onboarding and Offboarding

Customers begin by setting up their AWS environment including account creation and IAM configuration. The onboarding process involves creating Amazon S3 buckets for input and output data, establishing a HealthImaging data store, configuring IAM roles with appropriate permissions for import operations, and optionally installing the AWS CLI:

- DICOM P10 data is imported through S3 staging areas into HealthImaging data stores where it is automatically organized and optimized
- The service provides sample projects, SDKs, and comprehensive documentation to accelerate integration
- Customers can access HealthImaging through the AWS Management Console, AWS CLI, or programmatically via SDKs and APIs

Getting Started Guide:
https://docs.aws.amazon.com/healthimaging/latest/devguide/getting-started.html

Data Removal

Data removal in AWS HealthImaging involves deleting image sets and data stores through the console, CLI, or APIs. When customers delete data before the minimum 30-day retention period, additional charges apply for early deletion. The service provides clear data lifecycle management with the ability to delete specific image sets or entire data stores. Organizations can implement systematic data removal processes using AWS SDKs or APIs to ensure complete data cleanup during offboarding while maintaining compliance with healthcare data retention requirements.

Backup/Restore and Disaster Recovery

AWS HealthImaging provides inherent disaster recovery capabilities through AWS's globally distributed infrastructure and built-in data durability. The service automatically stores data across multiple Availability Zones within a region for high availability. HealthImaging leverages AWS's underlying storage infrastructure durability guarantees, and customers can implement additional disaster recovery strategies using cross-region replication patterns with AWS services. The service's cloud-native architecture provides resilience against infrastructure failures.

Backup Capabilities

AWS HealthImaging does not directly integrate with AWS Backup service. Instead, the service relies on AWS's underlying infrastructure for data durability and provides its own intelligent tiering system for data lifecycle management:

- Customers can implement backup strategies using the service's copy and export capabilities combined with S3 cross-region replication
- The service maintains data integrity through built-in pixel data verification and provides access to image sets across multiple availability zones for redundancy

Disaster Recovery

AWS HealthImaging does not directly integrate with AWS Elastic Disaster Recovery service. The service provides disaster recovery through AWS's multi-AZ architecture and regional availability across US East (N:

- Virginia), US West (Oregon), Asia Pacific (Sydney), and Europe (Ireland)
- Organizations can implement disaster recovery strategies using HealthImaging's APIs to replicate critical imaging data across regions and integrate with EventBridge for automated failover workflows

Recovery Objectives

AWS HealthImaging supports low RTO objectives through its high-availability architecture with subsecond image retrieval access from both Frequent Access and Archive Instant Access tiers. The service's multi-AZ deployment and progressive resolution capabilities

enable rapid recovery of imaging data. For RPO objectives, organizations can leverage HealthImaging's real-time import capabilities and EventBridge integration to maintain minimal data loss windows through automated replication and backup processes.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/healthimaging/latest/devguide/endpoints-quotas.html

FAQ: https://docs.aws.amazon.com/healthimaging/latest/devguide/what-is.html

Technical Requirements

Service Documentation:
https://docs.aws.amazon.com/healthimaging/latest/devguide/what-is.html

User Guide: https://docs.aws.amazon.com/healthimaging/latest/devguide/getting-started.html

API Reference:
https://docs.aws.amazon.com/healthimaging/latest/APIReference/index.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/healthimaging/pricing/

Cost Optimization

AWS HealthImaging provides unique cost optimization through intelligent tiering that automatically moves image sets between Frequent Access and Archive Instant Access tiers based on 30-day access patterns. The Archive Instant Access tier offers reduced storage costs while maintaining millisecond access latencies with no retrieval charges:

- The service eliminates traditional PACS infrastructure costs and provides pay-as-you-go pricing without upfront investments
- Additional optimization comes from HTJ2K compression efficiency, which reduces storage requirements while maintaining image quality, and indexed storage billing that separates metadata costs from image storage costs

Savings Considerations

Primary cost savings include elimination of on-premises PACS infrastructure, maintenance, and associated IT overhead. The intelligent tiering system automatically optimizes storage costs by moving infrequently accessed data to lower-cost tiers without manual intervention:

- Organizations save on bandwidth and compute costs through efficient HTJ2K compression and progressive resolution capabilities
- The service's cloud-native architecture eliminates backup infrastructure costs while providing inherent durability
- Additional savings come from reduced development time through developer-friendly APIs, integration with existing AWS services, and elimination of separate disaster recovery infrastructure investments

Service Name

AWS HealthLake

Service Overview

AWS HealthLake is a HIPAA-eligible, serverless, and fully managed service that makes it easy to store, search, and analyze healthcare data at scale using the Fast Healthcare Interoperability Resources (FHIR) R4 specification. HealthLake enables healthcare organizations to ingest FHIR data from multiple sources, automatically extract clinical insights using natural language processing, and transform unstructured medical data into analytics-ready formats. It provides a centralized healthcare data lake that supports regulatory compliance requirements such as CMS and ONC rules while offering sub-second query latency and seamless scalability for building modern healthcare applications.

Key Use Cases

- Population Health Analytics: Analyze large volumes of healthcare data to identify trends, monitor disease outbreaks, and support value-based care initiatives
- Clinical Decision Support: Build AI-powered patient profiles and clinical applications that provide comprehensive patient histories and evidence-based insights
- Healthcare Data Integration: Unify clinical data across multiple health systems and electronic medical records using standardized FHIR formats
- Research and Clinical Trials: Support clinical trial participant screening, longitudinal patient data analysis, and research data management
- Health Information Exchange: Enable provincial or regional health information exchanges and cross-jurisdiction healthcare partnerships

Features

- **High-Performance FHIR Datastore**: Enterprise-grade FHIR R4 repository with sub-second latency, comprehensive security, and seamless data ingestion at scale
- **Integrated Natural Language Processing**: HIPAA-eligible NLP libraries that extract medical entities, conditions, medications, and procedures from unstructured text
- **Zero-ETL Analytics**: Direct SQL querying capabilities through Amazon Athena with automatic FHIR-to-SQL transformation at petabyte scale
- **FHIR API Infrastructure**: Comprehensive FHIR search capabilities and transactional FHIR server supporting standard FHIR operations
- **SMART on FHIR Authorization**: Support for Substitutable Medical Applications and Reusable Technologies authorization alongside AWS Signature Version 4

- **Real-time Subscriptions**: FHIR Subscriptions for monitoring and responding to data changes through Amazon EventBridge or REST-Hook endpoints

Value Proposition

AWS HealthLake eliminates the complexity of building and managing healthcare data infrastructure while ensuring regulatory compliance through HIPAA eligibility, HITRUST CSF certification, and SOC 2 compliance. Organizations benefit from automatic data standardization using medical nomenclatures like ICD-10-CM and RxNorm, built-in natural language processing for extracting insights from clinical notes, and zero-ETL analytics capabilities that enable immediate querying without extensive data engineering. The service offers enterprise-grade performance with sub-second latency, seamless scalability, and comprehensive integration with the broader AWS ecosystem, enabling healthcare organizations to focus on innovation rather than infrastructure management while accelerating time-to-market for healthcare applications.

Service Integration

AWS HealthLake deeply integrates with core AWS services to provide comprehensive healthcare data solutions. Amazon Comprehend Medical provides integrated NLP capabilities for extracting medical entities from unstructured text:

- Amazon Athena enables direct SQL querying of FHIR data through zero-ETL transformation
- Amazon S3 serves as the underlying storage for data import/export operations and backup
- AWS Lake Formation manages data catalog and access control
- Amazon EventBridge supports real-time FHIR subscriptions
- AWS Identity and Access Management controls access permissions
- Amazon CloudWatch provides monitoring and logging
- AWS CloudTrail enables audit trails
- AWS PrivateLink ensures secure connectivity
- The service also integrates with Amazon SageMaker for machine learning, Amazon QuickSight for dashboards, and AWS Lambda for custom processing workflows

Onboarding and Offboarding

Customers begin by setting up an AWS account and creating IAM users with appropriate permissions. The setup process involves creating Amazon S3 buckets for input and output data, establishing IAM service roles for HealthLake access, and creating a FHIR datastore through the AWS Management Console, CLI, or SDKs:

- Organizations can configure either AWS Signature Version 4 or SMART on FHIR authorization based on their requirements

- Data import begins by uploading FHIR R4 formatted files to S3 and initiating import jobs through the HealthLake console or APIs
- The service automatically processes and indexes the data, making it available for queries and analytics
- AWS provides comprehensive documentation, code examples, and tutorials to guide customers through the implementation process

Getting Started Guide: https://docs.aws.amazon.com/healthlake/latest/devguide/getting-started.html

Data Removal

AWS HealthLake follows FHIR specification requirements where deleted data is versioned rather than permanently removed, meaning deleted resources are hidden from analysis but remain in the service for compliance purposes. Customers can delete individual FHIR resources or entire datastores through the console, CLI, or APIs. When a datastore is deleted, all associated data and metadata are removed from the service. Organizations must ensure proper data backup before deletion if required for regulatory compliance. Customer-managed encryption keys can be revoked to render data inaccessible, and AWS provides tools for data export before offboarding.

Backup/Restore and Disaster Recovery

AWS HealthLake provides built-in data durability through AWS infrastructure with automatic replication within the region. The service supports data export capabilities to Amazon S3 for backup purposes at $0.19 per GB exported. Organizations can implement cross-region replication strategies by exporting data to S3 and using S3 Cross-Region Replication. The service maintains data integrity through encryption at rest using customer-managed keys and in-transit encryption. However, HealthLake does not provide native automated backup scheduling, requiring customers to implement custom backup strategies using export jobs and S3 lifecycle policies.

Backup Capabilities

AWS HealthLake does not natively integrate with AWS Backup service due to its specialized healthcare data format and FHIR compliance requirements. However, the service provides FHIR data export functionality that allows customers to export complete datasets to Amazon S3 in FHIR format for backup purposes:

- Organizations can implement automated backup strategies using AWS Lambda functions to schedule regular export jobs, combined with S3 lifecycle policies for long-term retention
- The exported data maintains FHIR R4 compliance and can be re-imported into new HealthLake datastores if needed

Disaster Recovery

AWS HealthLake does not directly integrate with AWS Elastic Disaster Recovery due to its unique FHIR data format and managed service architecture. However, organizations can implement disaster recovery strategies using multi-region architectures with data export/import capabilities:

- The service supports exporting complete datasets to S3, which can be replicated across regions using S3 Cross-Region Replication
- In disaster scenarios, new HealthLake datastores can be created in alternate regions and populated using previously exported FHIR data
- Organizations typically implement application-level disaster recovery patterns rather than infrastructure-level replication

Recovery Objectives

AWS HealthLake supports flexible RPO and RTO objectives through its export and import capabilities. Organizations can achieve RPO targets by scheduling regular data exports based on their data change frequency requirements. RTO objectives depend on data volume and import speed, with HealthLake supporting concurrent import jobs to accelerate recovery. For lower RTO requirements, organizations can implement multi-region active-passive architectures with regular data synchronization. The service's managed nature and automatic scaling help minimize recovery complexity, though customers must design appropriate backup and recovery workflows based on their specific healthcare compliance and operational requirements.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/healthlake/latest/devguide/reference-healthlake-endpoints-quotas.html

FAQ: https://aws.amazon.com/healthlake/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/healthlake/latest/devguide/

User Guide: https://docs.aws.amazon.com/healthlake/latest/devguide/

API Reference: https://docs.aws.amazon.com/healthlake/latest/APIReference/

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/healthlake/pricing/

Cost Optimization

AWS HealthLake offers tiered pricing with Advanced ($0.27/hour) and Standard ($0.25/hour) options, allowing organizations to select appropriate feature sets. Cost optimization strategies include choosing the right tier based on required features, optimizing query patterns to stay within the included 3,500 queries per hour limit, and managing data export frequency since exports cost $0.19 per GB:

- Organizations can optimize NLP costs at $0.0010 per 100 characters by preprocessing data to remove unnecessary text
- FHIR subscription costs can be managed by selecting appropriate delivery methods (EventBridge at $2.00 vs REST-Hook at $0.90 per million events)
- Storage costs scale with data volume, so implementing data lifecycle policies and archiving older records can provide savings

Savings Considerations

Primary cost savings opportunities include selecting the Standard tier when advanced features aren't required, reducing potential costs by up to $0.02 per hour per datastore. Organizations can achieve significant savings by optimizing data export strategies, implementing scheduled exports during off-peak periods, and using S3 lifecycle policies for long-term storage:

- Query optimization can prevent overage charges beyond the included 3,500 queries per hour limit
- NLP cost management through data preprocessing and selective text analysis can reduce character-based charges
- Implementing efficient FHIR subscription patterns and choosing cost-effective delivery methods helps control event-based pricing
- Regional selection based on data residency requirements and proximity can optimize data transfer costs while maintaining compliance with healthcare regulations

Service Name

AWS HealthOmics

Service Overview

AWS HealthOmics is a HIPAA-eligible service that accelerates clinical diagnostic testing, drug discovery, and agricultural research by fully managing the complex infrastructure behind bioinformatics workflows. The service supports industry-standard workflow languages (WDL, Nextflow, CWL) and seamlessly scales bioinformatics infrastructure to support data from tens of thousands of tests per day with predictable cost-per-sample. HealthOmics consists of three main components: workflows for processing and analyzing genomics data, storage for genomics data and genome references, and analytics for transforming and analyzing genomics data using variant and annotation stores.

Key Use Cases

- Clinical Diagnostics: Build and scale diagnostic testing workflows with predictable costs and fully managed infrastructure that grows with testing volume, supporting variant calling, filtering, and annotation
- Drug Discovery: Accelerate therapeutic research by orchestrating biological foundation models at scale, enabling rapid iteration across millions of potential candidates using AI-powered analysis
- Agricultural Research: Enhance crop traits like drought tolerance and pest resistance through AI-powered workflows that improve food security and agricultural productivity using genomic analysis

Features

- **Workflows**: Process and analyze genomics data using WDL, Nextflow, or CWL with private and Ready2Run pre-built workflows
- **Storage**: Purpose-built storage for bioinformatics file formats with sequence stores for genomics data and reference stores for genome references
- **Analytics**: Transform and analyze genomics data using variant stores and annotation stores with population-scale data support
- **Auto-scaling**: Seamlessly scales across 100,000+ concurrent vCPUs to support tens of thousands of tests daily
- **Data Collaboration**: Secure sharing capabilities with fine-grained access controls, tagging, and permissions
- **HIPAA Compliance**: Built-in compliance features with comprehensive audit trails, data provenance tracking, and encryption

Value Proposition

AWS HealthOmics eliminates the complexity of managing bioinformatics infrastructure while providing predictable cost-per-sample pricing and enterprise-scale performance. The service handles technical complexities like managing compute resources and maintaining workflow engines, allowing organizations to focus entirely on scientific breakthroughs. Key differentiators include: fully managed infrastructure that scales to 100,000+ concurrent vCPUs, support for industry-standard workflow languages, HIPAA-eligible infrastructure with built-in compliance features, comprehensive data provenance and audit trails, seamless integration with AWS services ecosystem, and predictable pricing with no upfront costs.

Service Integration

AWS HealthOmics integrates extensively with core AWS services to provide a comprehensive bioinformatics platform. Primary integrations include Amazon S3 for data storage and input/output operations, Amazon Elastic Container Registry (ECR) for workflow container storage, AWS Lake Formation and Amazon Athena for analytics and querying genomics data through Apache Iceberg tables, Amazon SageMaker AI for machine learning workflows, Amazon EventBridge for event-driven architectures and workflow automation, AWS Lambda for custom processing, Amazon CloudWatch for monitoring and logging, and AWS IAM for access control and security management.

Onboarding and Offboarding

Getting started with AWS HealthOmics requires proper IAM permissions and roles for the service. Customers can begin by using Ready2Run workflows in the HealthOmics console, selecting pre-built workflows, entering run names and output locations, and creating service roles:

- The service supports Amazon Q CLI for natural language workflow management
- Setup involves creating AWS accounts, securing root users with MFA, creating administrative users with IAM Identity Center, configuring IAM permissions for HealthOmics, and optionally connecting with external code repositories through AWS CodeConnections for GitHub integration

Getting Started Guide: https://docs.aws.amazon.com/omics/latest/dev/getting-started.html

Data Removal

Data removal in AWS HealthOmics involves deleting resources through the console, CLI, or APIs. Customers can delete sequence stores, reference stores, workflows, variant stores, and annotation stores. Data in storage follows minimum retention periods of 30 days with prorated charges for early deletion. Cross-account sharing capabilities require

proper permission management during offboarding to ensure data access is properly revoked from shared accounts.

Backup/Restore and Disaster Recovery

AWS HealthOmics provides data durability through its purpose-built storage architecture with automatic data integrity verification using ETags. The service maintains data relationships for provenance tracking and supports cross-region capabilities. Data in ACTIVE state is readily accessible, while data not accessed for 30+ days moves to ARCHIVE state, requiring reactivation for access.

Backup Capabilities

AWS HealthOmics does not explicitly integrate with AWS Backup service. Instead, it provides built-in data durability through its storage architecture with automatic ETag generation for content integrity validation. The service preserves data relationships for provenance and supports data export capabilities to Amazon S3 for external backup strategies if needed.

Disaster Recovery

AWS HealthOmics does not directly integrate with AWS Elastic Disaster Recovery. The service provides multi-region availability and supports cross-region data operations. Disaster recovery is primarily achieved through the service's distributed architecture across AWS availability zones and the ability to export data to S3 for cross-region replication strategies.

Recovery Objectives

AWS HealthOmics supports recovery objectives through its multi-region availability, automatic data integrity verification, and export capabilities to S3. The service's distributed architecture helps achieve faster recovery times, while data export features enable customers to implement cross-region backup strategies to meet specific RPO requirements based on their genomics data criticality.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/general/latest/gr/healthomics-quotas.html

FAQ: https://aws.amazon.com/healthomics/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/omics/latest/dev/index.html

User Guide: https://docs.aws.amazon.com/omics/latest/dev/what-is-healthomics.html

API Reference: https://docs.aws.amazon.com/omics/latest/api/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/healthomics/pricing/

Cost Optimization

AWS HealthOmics offers unique cost optimization through predictable cost-per-sample pricing and automated storage tiering. The service provides a Free Tier with 1500 gigabase-months in active/archive storage, 275 omics.m.xlarge instance hours, and 49,000 GB-hours of run storage. Cost optimization strategies include using archive storage for infrequently accessed data, leveraging Ready2Run workflows for predictable pricing versus private workflows, utilizing run caching to avoid recomputing identical tasks, implementing proper storage lifecycle management with 30-day minimum retention, and optimizing workflow resource requests to match actual computational needs.

Savings Considerations

Primary cost savings opportunities include leveraging the generous Free Tier allowances for initial workloads, using archive storage class for long-term data retention at reduced costs, implementing run caching strategies to eliminate redundant computations, choosing appropriate instance types and storage capacities based on actual workflow requirements, utilizing Ready2Run workflows which offer predictable per-run pricing versus variable private workflow costs, optimizing data formats and compression to reduce storage costs, and implementing proper data lifecycle policies to automatically transition data to lower-cost storage tiers while maintaining compliance with minimum retention requirements.

Service Name

AWS Health Dashboard

Service Overview

AWS Health Dashboard is a service that provides customers with ongoing visibility into the performance and availability of their AWS resources and services. It offers real-time alerts and notifications for health changes, enabling users to manage events and prepare for planned activities. The dashboard displays both account-specific events that affect your resources and public events that are general service issues. Users can view open, recent, and scheduled changes, as well as notifications and event logs from the past 90 days. The service integrates with Amazon EventBridge for automation and AWS User Notifications for centralized notification management.

Key Use Cases

- Use case 1: Proactive incident management by monitoring AWS Health events to identify potential issues affecting applications before they impact users
- Use case 2: Scheduled maintenance planning by viewing upcoming AWS service changes and maintenance windows to prepare applications and inform stakeholders
- Use case 3: Multi-account organizational health monitoring using organizational view to aggregate health events across all AWS accounts in an organization for centralized visibility

Features

- **Account Events Dashboard**: Displays events specific to your AWS account including open, recent, and scheduled changes with detailed event information
- **Public Events View**: Shows general AWS service issues and interruptions that are not specific to individual accounts
- **Calendar View**: Provides monthly calendar format for viewing AWS Health events and scheduled changes
- **Event Filtering**: Filter events by status, start time, region, service, and other parameters for focused monitoring
- **Resource Export**: Download affected resources list in CSV or JSON format for external analysis
- **EventBridge Integration**: Automatically trigger actions and notifications through Amazon EventBridge rules
- **Organizational View**: Centralized view of health events across all accounts in AWS Organizations

- **Timeline View**: Historical view of events and issues that have affected resources over time

Value Proposition

AWS Health Dashboard is accessible to all customers at no additional cost and requires no setup, providing immediate value through its console interface. Unlike third-party monitoring solutions, it provides authoritative information directly from AWS about service health and account-specific events:

- The service offers deep integration with AWS services through EventBridge for automation and can aggregate events across organizational boundaries
- Its real-time nature and comprehensive event categorization (account notifications, issues, and scheduled changes) help accelerate troubleshooting and maintain optimal AWS usage, making it superior to manual monitoring or generic health check services

Service Integration

AWS Health Dashboard integrates deeply with Amazon EventBridge to enable automated responses to health events, allowing customers to trigger Lambda functions, send notifications via SNS, or create incident management workflows. It works with AWS Organizations to provide organizational view across multiple accounts and integrates with AWS User Notifications for centralized notification management:

- The service also connects with Amazon CloudWatch for enhanced monitoring capabilities and can integrate with JIRA and ServiceNow through the Service Management Connector
- Additionally, it supports integration with chat applications like Amazon Chime and Slack for real-time team notifications

Onboarding and Offboarding

Getting started with AWS Health Dashboard requires only an AWS account as the service is available to all customers at no additional cost with no setup required. Simply sign in to the AWS Management Console and navigate to the AWS Health Dashboard to view events:

- For enhanced functionality, customers can configure Amazon EventBridge rules to automate responses to health events and enable organizational view for multi-account visibility
- Advanced features like the AWS Health API require Business, Enterprise On-Ramp, or Enterprise Support plans
- The dashboard is immediately accessible and provides 90 days of event history upon first access

Getting Started Guide: https://docs.aws.amazon.com/health/latest/ug/getting-started-health-dashboard.html

Data Removal

AWS Health Dashboard operates as a read-only service that displays health events and does not store customer data that requires removal. Events are automatically purged after 90 days as part of the standard retention policy. No explicit offboarding process is required since the service displays real-time AWS service status information rather than storing persistent customer data that needs deletion.

Backup/Restore and Disaster Recovery

AWS Health Dashboard does not provide backup and disaster recovery capabilities as it is a monitoring and notification service that displays real-time health information. The service itself is highly available across AWS regions and automatically delivers events to EventBridge. Customers can export affected resources data in CSV or JSON format for external backup purposes if needed for compliance or audit requirements.

Backup Capabilities

AWS Health Dashboard does not have traditional backup capabilities and does not integrate with AWS Backup since it is a monitoring service that displays real-time health events rather than storing persistent data. The service maintains 90 days of event history which serves as historical reference rather than backup data.

Disaster Recovery

AWS Health Dashboard does not directly support disaster recovery operations through AWS Elastic Disaster Recovery. However, it can be used as part of disaster recovery planning by monitoring health events and integrating with EventBridge to trigger automated disaster recovery responses when specific health events occur.

Recovery Objectives

AWS Health Dashboard helps customers meet RPO and RTO objectives by providing real-time visibility into AWS service health issues and planned maintenance events. Through EventBridge integration, customers can automate incident response workflows to reduce RTO. The service enables proactive planning for scheduled changes to minimize service disruption and optimize recovery time objectives.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/health/latest/ug/index.html

FAQ: https://docs.aws.amazon.com/health/latest/ug/what-is-aws-health.html

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/health/latest/ug/index.html

User Guide: https://docs.aws.amazon.com/health/latest/ug/getting-started-health-dashboard.html

API Reference: https://docs.aws.amazon.com/health/latest/APIReference/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://docs.aws.amazon.com/health/latest/ug/what-is-aws-health.html

Cost Optimization

AWS Health Dashboard is provided at no additional cost to all AWS customers, making it a cost-effective monitoring solution. The only potential costs come from downstream services triggered by EventBridge integrations, such as Lambda functions, SNS notifications, or SQS queues:

- Customers can optimize costs by implementing efficient EventBridge rules that filter events appropriately to avoid unnecessary triggering of downstream services
- The organizational view feature provides cost savings by centralizing monitoring across multiple accounts without additional charges

Savings Considerations

The primary cost savings come from the service being completely free with no setup costs or ongoing charges. By providing early warning of issues and scheduled maintenance, it can prevent costly downtime and enable proactive resource management:

- Integration with automation through EventBridge can reduce operational overhead and manual monitoring costs
- For organizations with multiple AWS accounts, the organizational view eliminates the need for separate monitoring tools or manual aggregation processes, providing significant operational cost savings through centralized health monitoring

Service Name

AWS Identity and Access Management (IAM)

Service Overview

AWS Identity and Access Management (IAM) is a web service that helps securely control access to AWS resources by managing permissions for users and entities. IAM controls who is authenticated (signed in) and authorized (has permissions) to use AWS resources. When an AWS account is created, it includes a root user with complete access to all services, but IAM enables creation of additional identities with specific permissions. IAM provides the infrastructure necessary to control authentication and authorization across AWS accounts, offering centralized identity management, fine-grained access control, and security features like multi-factor authentication and identity federation for secure resource access.

Key Use Cases

- Enterprise user management: Connect corporate Active Directory to IAM Identity Center for centralized access control across multiple AWS accounts and applications
- Application security: Grant applications running on EC2 instances secure access to AWS services using IAM roles with temporary credentials
- Cross-account access: Enable secure access between different AWS accounts for partners, vendors, or organizational divisions using cross-account roles
- Compliance and governance: Implement least-privilege access controls, audit trails, and permissions boundaries to meet regulatory requirements
- Multi-factor authentication: Add extra security layer requiring users to provide codes from MFA devices in addition to passwords

Features

- **IAM Users and Groups**: Create individual users and organize them into groups with specific permissions for single persons or applications
- **IAM Roles**: Provide temporary credentials for federated access, cross-account access, and applications running on EC2
- **Policies and Permissions**: Define granular permissions using JSON policy language with identity-based and resource-based policies
- **Multi-Factor Authentication (MFA)**: Support for passkeys, security keys, virtual authenticator apps, and hardware TOTP tokens for enhanced security
- **Identity Federation**: Enable users with existing passwords from corporate networks or identity providers to access AWS using temporary credentials

- **IAM Access Analyzer**: Generate least-privilege policies, verify public access, and validate IAM policies for secure permissions
- **Permissions Boundaries**: Set maximum permissions for identity-based policies to establish guardrails for delegated administration
- **Service-Linked Roles**: Predefined IAM roles created automatically when AWS services need permissions to access resources on your behalf

Value Proposition

IAM provides centralized identity and access management with no additional charges, integrating seamlessly with over 250 AWS services. Unlike third-party solutions, IAM offers native integration with AWS services, automatic scaling, and built-in compliance features including PCI DSS validation:

- IAM enables fine-grained access control through JSON policies, supports multiple authentication methods including MFA and federation, and provides eventual consistency with high availability through multi-server replication
- Organizations benefit from reduced complexity in managing access across multiple AWS accounts, enhanced security through least-privilege principles, and comprehensive audit capabilities through CloudTrail integration

Service Integration

IAM integrates with virtually all AWS services as the foundational security service for access management. Core integrations include AWS CloudTrail for detailed logging and auditing, AWS Organizations for multi-account management, AWS STS for temporary credential generation, and AWS Identity Center for centralized user access:

- IAM works with compute services like EC2 for instance roles, storage services like S3 for bucket policies, and management services like CloudFormation for infrastructure-as-code deployments
- The service also integrates with AWS SDKs and CLI tools for programmatic access management, ensuring consistent security controls across all AWS service interactions

Onboarding and Offboarding

Getting started with IAM requires an existing AWS account and begins with securing the root user by enabling multi-factor authentication. Create an administrative user for everyday tasks rather than using root credentials:

- Set up IAM users for individuals or applications, organize them into groups for easier permission management, and attach appropriate AWS managed policies
- Configure IAM roles for applications running on EC2 instances or for cross-account access scenarios

- Implement MFA for enhanced security and use IAM Access Analyzer to validate policies
- Follow the principle of least privilege by starting with minimal permissions and adding access as needed

Getting Started Guide: https://docs.aws.amazon.com/serverless/latest/devguide/starter-iam.html

Data Removal

IAM data removal involves systematically deleting IAM identities and associated resources. Remove IAM users by first deleting their access keys, detaching policies, removing from groups, and finally deleting the user account. Delete IAM roles by first detaching policies and removing trust relationships. Clean up unused policies, groups, and service-linked roles. Use IAM Access Analyzer to identify unused access before removal. IAM changes are eventually consistent, so allow time for propagation across AWS regions.

Backup/Restore and Disaster Recovery

IAM configuration itself does not require traditional backup as it is a managed service with built-in high availability and eventual consistency through multi-server replication across regions. However, organizations should document IAM policies, roles, and user configurations as part of infrastructure-as-code practices using CloudFormation or Terraform. IAM changes are automatically replicated across AWS regions, providing inherent disaster recovery capabilities for identity and access management configurations.

Backup Capabilities

IAM does not provide traditional backup capabilities as it is a managed service with built-in redundancy and replication. IAM configurations are automatically maintained across multiple servers for high availability:

- Organizations should implement infrastructure-as-code practices to version control and recreate IAM configurations
- IAM does not integrate with AWS Backup service as the identity management data is inherently protected through AWS's managed service architecture

Disaster Recovery

IAM inherently supports disaster recovery through its eventually consistent, globally replicated architecture across multiple servers and regions. IAM does not integrate with AWS Elastic Disaster Recovery as it is a foundational service with built-in resilience:

- During regional failures, IAM continues operating from other regions, though some changes may experience temporary delays during propagation
- Cross-region IAM role assumptions and federated access continue functioning during regional outages

Recovery Objectives

IAM supports recovery objectives through its highly available, eventually consistent architecture that automatically replicates data across multiple servers. While IAM changes may take time to propagate (eventual consistency), the service maintains continuous availability during regional failures. Organizations achieve low RTO by implementing cross-region IAM roles and federation strategies. RPO is effectively zero for IAM configurations due to automatic replication, though policy changes may experience brief propagation delays.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/IAM/latest/UserGuide/reference_iam-quotas.html

FAQ: https://aws.amazon.com/iam/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/IAM/latest/UserGuide/introduction.html

User Guide: https://docs.aws.amazon.com/IAM/latest/UserGuide/introduction.html

API Reference: https://docs.aws.amazon.com/IAM/latest/APIReference/welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/iam/pricing/

Cost Optimization

IAM itself incurs no additional charges as it is provided free with all AWS accounts. Cost optimization focuses on managing the resources that IAM-authenticated users and roles access rather than IAM service costs:

- Use IAM policies to implement resource-based restrictions that prevent unnecessary resource provisioning or access to expensive services
- Implement permissions boundaries to limit maximum spending capabilities of delegated administrators
- Use IAM roles instead of long-term access keys to reduce security overhead and potential unauthorized usage costs
- Regular access reviews through IAM Access Analyzer help identify and remove unused permissions that could lead to unintended resource usage

Savings Considerations

Primary cost savings come from preventing unauthorized or accidental resource usage through proper IAM controls rather than direct IAM service costs. Implement least-privilege policies to prevent users from provisioning expensive resources unnecessarily:

- Use IAM conditions in policies to restrict actions based on time, IP address, or other factors to limit when expensive operations can occur
- Regular access reviews and cleanup of unused IAM users, roles, and permissions reduce potential for unauthorized spending
- Cross-account roles eliminate need for duplicate resources across accounts
- Proper IAM governance prevents shadow IT spending and ensures resources are only accessible to authorized personnel for legitimate business purposes

Service Name

AWS IoT Core

Service Overview

AWS IoT Core is a managed cloud service that enables connected devices to securely interact with cloud applications and other devices. It acts as a central hub for bi-directional communication between IoT devices and AWS services, supporting billions of devices and trillions of messages. The service provides device gateway capabilities, message broker functionality, and secure device management features including authentication, authorization, and encrypted communication using TLS protocols. AWS IoT Core processes and routes messages to AWS endpoints and other devices reliably and securely while offering features like device shadows for offline state management and rules engine for data processing and routing.

Key Use Cases

- Connected home automation: Smart devices, appliances, security systems, and voice-enabled products with Alexa integration for home automation and monitoring
- Industrial IoT and predictive maintenance: Asset condition monitoring, predictive quality analytics, equipment performance optimization, and process automation for manufacturing and industrial environments
- Vehicle connectivity and fleet management: Connected mobility solutions for automotive applications, vehicle data collection, telematics, and fleet monitoring with real-time analytics

Features

- **Device Gateway**: Secure bi-directional communication bridge between IoT devices and AWS IoT Core supporting MQTT, HTTP, and WebSocket protocols
- **Message Broker**: Scalable messaging service for real-time communication between devices and AWS services with MQTT protocol support
- **Device Shadow**: Virtual representation of devices maintaining state information for online/offline communication and synchronization
- **Rules Engine**: SQL-like service for filtering, transforming, and routing device data to other AWS services and external endpoints
- **Thing Registry**: Device metadata database storing certificates, attributes, and thing types for organized device management
- **Authentication and Authorization**: Fine-grained security with X.509 certificates, IAM integration, and IoT policies for secure device access
- **LoRaWAN Support**: Managed LoRaWAN Network Server for low-power, long-range connectivity with public and private network support

- **Device Advisor**: Testing service for validating IoT devices before deployment with compatibility and capability assessments

Value Proposition

AWS IoT Core offers unmatched scalability supporting billions of devices and trillions of messages with global availability across 20+ regions. The service provides enterprise-grade security with end-to-end encryption, certificate-based authentication, and fine-grained access controls:

- Its fully managed nature eliminates infrastructure management overhead while offering seamless integration with 20+ AWS services including Lambda, S3, DynamoDB, Kinesis, and analytics services
- The pay-as-you-use pricing model with generous free tier makes it cost-effective for any scale
- Advanced features like message enrichment, device shadows for offline scenarios, and comprehensive SDKs across multiple languages accelerate development while ensuring reliability and performance

Service Integration

AWS IoT Core integrates natively with over 20 AWS services through its Rules Engine. Primary integrations include AWS Lambda for serverless computing, Amazon S3 for data storage, Amazon DynamoDB for NoSQL database operations, and Amazon Kinesis for real-time data streaming:

- Additional integrations encompass Amazon SNS for notifications, Amazon SQS for message queuing, AWS Step Functions for workflow automation, Amazon CloudWatch for monitoring, AWS IoT Analytics for data analysis, Amazon Timestream for time-series data, and Amazon API Gateway for REST APIs
- The service also connects to Amazon Elasticsearch, AWS IoT Events for event detection, and supports external HTTP endpoints and Apache Kafka clusters for hybrid architectures

Onboarding and Offboarding

Customers begin by setting up an AWS account and accessing the AWS IoT console. The onboarding process involves creating a ‘thing’ object to represent the device, generating and downloading security certificates and keys, configuring device policies for access permissions, and installing AWS IoT Device SDKs on target devices:

- AWS provides multiple getting-started paths including quick connect tutorials for immediate hands-on experience, interactive demos requiring no physical devices, and comprehensive developer tutorials with detailed explanations
- Device Advisor helps validate device compatibility before production deployment

- The service offers extensive documentation, code examples across multiple programming languages, and SDKs supporting C++, Python, JavaScript, Java, and embedded systems

Getting Started Guide: https://docs.aws.amazon.com/iot/latest/developerguide/iot-gs.html

Data Removal

Offboarding requires systematic removal of IoT resources including deleting thing objects, revoking and deleting device certificates, removing device policies and thing groups, and clearing device shadow data. Customers should disable rules engine configurations, delete custom authorizers, and remove IoT jobs before account closure. All data stored in integrated AWS services like S3 buckets, DynamoDB tables, and CloudWatch logs must be separately deleted as these are not automatically removed when IoT Core resources are deleted.

Backup/Restore and Disaster Recovery

AWS IoT Core provides disaster recovery through cross-region replication of device registry, certificates, and policies using DynamoDB Global Tables and Lambda functions. Device shadows and thing metadata can be automatically replicated to secondary regions. AWS offers a dedicated Disaster Recovery for AWS IoT solution using Step Functions workflows for automated failover. The service leverages Route 53 health checks for automatic region failover and supports backup strategies through AWS Backup integration for connected services.

Backup Capabilities

AWS IoT Core does not directly integrate with AWS Backup service. Backup capabilities exist through replication of device registry data, certificates, and policies to secondary regions using DynamoDB streams and Lambda functions:

- Device shadow state and thing metadata can be backed up through custom solutions
- Connected AWS services like S3, DynamoDB, and CloudWatch that store IoT data can utilize AWS Backup for comprehensive data protection strategies

Disaster Recovery

AWS IoT Core does not directly integrate with AWS Elastic Disaster Recovery service. However, it provides disaster recovery capabilities through multi-region architecture, cross-region replication of device registry and certificates, and automated failover using Route 53 health checks. AWS offers a dedicated IoT disaster recovery solution using Step Functions and DynamoDB Global Tables for real-time replication and automated recovery procedures.

Recovery Objectives

AWS IoT Core supports RPO and RTO objectives through multi-region deployment strategies. Cross-region replication of device registry and certificates provides near real-time data protection with RPO of minutes. Automated failover using Route 53 and health checks can achieve RTO of 2-4 hours for backup/restore strategies. For critical workloads, warm standby or multi-site active/active architectures can deliver sub-hour RTO requirements while maintaining low RPO through continuous replication.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/general/latest/gr/iot-core.html

FAQ: https://aws.amazon.com/iot-core/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/iot/latest/developerguide/what-is-aws-iot.html

User Guide: https://docs.aws.amazon.com/iot/latest/developerguide/iot-gs.html

API Reference: https://docs.aws.amazon.com/iot/latest/apireference/

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/iot-core/pricing/

Cost Optimization

AWS IoT Core offers multiple cost optimization strategies including efficient message batching to reduce per-message costs, Basic Ingest feature for free message ingestion when using reserved topics, and payload compression to minimize data transfer charges. Device connection optimization through persistent sessions reduces reconnection overhead:

- Strategic use of device shadows minimizes frequent state updates, while Rules Engine filtering reduces downstream service calls
- The free tier provides 2.25M connection minutes, 500K messages, and 225K registry operations monthly
- Organizations can leverage LoRaWAN for low-cost, low-power connectivity and implement edge processing with IoT Greengrass to reduce cloud data transfer costs

Savings Considerations

Primary cost savings opportunities include utilizing the 12-month free tier effectively, implementing message compression and batching strategies, and optimizing device connection patterns to minimize connection time charges. Basic Ingest provides free message processing when using specific topic formats:

- Edge computing with AWS IoT Greengrass reduces cloud processing costs by handling data locally
- Strategic data retention policies and lifecycle management for stored IoT data in S3 and DynamoDB minimize long-term storage costs
- Monitoring usage patterns through Cost Explorer helps identify optimization opportunities, while right-sizing concurrent connections and message frequencies based on actual requirements prevents over-provisioning charges

Service Name

AWS IoT Device Management

Service Overview

AWS IoT Device Management is a cloud-based service that enables secure management of IoT devices at scale throughout their lifecycle. It integrates with AWS IoT Core to connect devices to the cloud and other devices for remote management. The service helps customers onboard, organize, monitor, and remotely update IoT devices ranging from simple sensors to complex systems. It provides capabilities for bulk device registration, fleet indexing and search, device organization into hierarchical groups, remote messaging and commands, over-the-air software updates, secure tunneling for troubleshooting, and comprehensive device monitoring with fine-grained logging.

Key Use Cases

- Fleet Management: Organize and manage large fleets of IoT devices with hierarchical grouping, bulk operations, and automated device lifecycle management
- Remote Device Operations: Send commands, retrieve device logs, change device states, and perform remote diagnostics on devices in the field
- Over-the-Air Updates: Push software and firmware updates to devices with controlled rollout schedules, failure thresholds, and rollback capabilities
- Device Troubleshooting: Establish secure tunneling connections to devices behind firewalls for remote validation, diagnosis, and corrective actions
- Device Monitoring and Analytics: Track device performance, connectivity status, and operational metrics with fleet indexing and search capabilities

Features

- **Bulk Registration**: Register multiple devices using templates with device information, certificates, and configurations
- **Fleet Indexing and Search**: Query device records and aggregate statistics based on attributes, states, and connectivity
- **Device Jobs**: Run and monitor software updates and remote operations on single devices or entire fleets
- **Secure Tunneling**: Create secure communications sessions to individual devices for troubleshooting without firewall changes
- **Device Commands**: Send low-latency remote commands to devices with real-time execution status tracking
- **Software Package Management**: Track and monitor software package versions with centralized dashboard and targeted updates

- **Device Organization**: Group devices hierarchically based on function, security requirements, or other categories
- **Managed Integrations**: Pre-built integrations for ZigBee, Z-Wave, and Wi-Fi devices with unified device control

Value Proposition

AWS IoT Device Management provides comprehensive device lifecycle management at scale with built-in security, reliability, and integration with the broader AWS ecosystem. Unlike building custom solutions, it offers managed services that automatically handle device onboarding, secure communications, and fleet operations without infrastructure overhead:

- The service provides enterprise-grade security with granular access controls, encrypted communications, and secure tunneling capabilities
- Its tight integration with AWS IoT Core enables seamless device-to-cloud connectivity, while integration with 200+ AWS services allows for advanced analytics, machine learning, and business logic implementation
- The pay-as-you-use pricing model eliminates upfront costs, and the 99.9% SLA ensures reliable operations for business-critical IoT deployments

Service Integration

AWS IoT Device Management tightly integrates with AWS IoT Core as the foundation for device connectivity and messaging. It works seamlessly with AWS IoT Device Defender for continuous security monitoring and threat detection:

- Integration with AWS Lambda enables custom business logic and event processing for device operations
- Amazon DynamoDB and Amazon S3 provide scalable storage for device data and firmware packages
- AWS IAM manages fine-grained access control and permissions for device operations
- AWS CloudWatch provides monitoring, logging, and alerting for device fleets
- Amazon API Gateway enables RESTful APIs for device management applications
- AWS Step Functions orchestrates complex device provisioning and update workflows
- The service also integrates with AWS Backup for data protection and Amazon Route 53 for disaster recovery scenarios

Onboarding and Offboarding

Customers begin by accessing the AWS IoT Device Management console and creating their first IoT things (device representations) with associated certificates and policies. The bulk

registration feature allows uploading templates containing device information for large-scale deployments:

- Device grouping enables organizing devices hierarchically based on function or security requirements
- Fleet indexing must be enabled to search and query device attributes and states
- For managed integrations, customers configure hub devices and use simple setup, zero-touch setup, or user-guided setup flows depending on their device types
- Prerequisites include AWS CLI installation, appropriate IAM permissions, and device SDK integration
- The service provides step-by-step tutorials and sample code for common scenarios to accelerate implementation

Getting Started Guide: https://docs.aws.amazon.com/iot/latest/developerguide/iot-gs.html

Data Removal

Device offboarding involves deleting device things, certificates, and policies from the AWS IoT registry using the delete-managed-thing command or console interface. Device shadow data and historical logs can be removed through service APIs. For managed integrations, devices are removed from hub associations and credential lockers are deleted. Customers should ensure all device jobs are completed or cancelled before removal. Data stored in integrated services like DynamoDB or S3 requires separate deletion processes following each service's data removal procedures.

Backup/Restore and Disaster Recovery

AWS IoT Device Management provides disaster recovery capabilities through cross-region replication of device registries, certificates, and policies using AWS solutions. Device shadow data and configurations can be replicated to secondary regions using DynamoDB global tables and AWS Lambda functions. The service integrates with AWS Backup for protecting associated data in other AWS services. Amazon Route 53 health checks and traffic policies enable automatic failover to secondary regions during disasters.

Backup Capabilities

The service does not have native backup capabilities but relies on cross-region replication strategies. Device registry information, certificates, and policies can be replicated using AWS Lambda and DynamoDB streams:

- AWS Backup can protect associated data stored in integrated services like S3 and DynamoDB
- Customers must implement custom backup strategies for device configurations and fleet settings using AWS infrastructure-as-code practices

Disaster Recovery

AWS IoT Device Management supports disaster recovery through AWS solutions that replicate device shadows, certificates, and policies to secondary regions. The service does not directly integrate with AWS Elastic Disaster Recovery but provides cross-region replication capabilities using DynamoDB global tables and Lambda functions. Amazon Route 53 enables automatic failover between regions based on health checks of the MQTT message broker.

Recovery Objectives

The service helps meet RPO objectives through real-time replication of device registry data and configurations to secondary regions. RTO objectives can be achieved through automated failover mechanisms using Amazon Route 53 and pre-provisioned resources in disaster recovery regions. Cross-region replication typically provides RPO of minutes and RTO of 2-4 hours depending on implementation complexity and failover automation level.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/general/latest/gr/iot_device_management.html

FAQ: https://aws.amazon.com/iot-device-management/faq/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/iot/latest/developerguide/what-is-aws-iot.html

User Guide: https://docs.aws.amazon.com/iot/latest/developerguide/iot-thing-management.html

API Reference: https://docs.aws.amazon.com/iot/latest/developerguide/iot-remote-command.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/iot-device-management/pricing/

Cost Optimization

AWS IoT Device Management offers several cost optimization strategies including usage-based pricing with no minimum fees or upfront costs. Customers can optimize costs by batching device operations, using efficient data transfer protocols, and implementing device data compression:

- Fleet indexing costs can be controlled by limiting custom fields and query complexity
- Device jobs can be optimized through targeted deployments rather than fleet-wide operations
- Secure tunneling costs are minimized by limiting session duration and concurrent connections
- The service provides detailed pricing metrics for bulk registration, device jobs, and secure tunneling to enable precise cost tracking and optimization

Savings Considerations

Primary cost savings come from eliminating the need to build and maintain custom device management infrastructure. The managed service model reduces operational overhead, development time, and infrastructure costs compared to self-hosted solutions:

- Pay-per-use pricing eliminates over-provisioning costs common with traditional infrastructure
- Integration with AWS services enables consolidated billing and volume discounts
- Efficient device grouping and bulk operations reduce per-device management costs
- Automated device lifecycle management reduces manual operational expenses
- The 12-month free tier provides 50 remote actions per month, enabling cost-effective proof-of-concept development and small-scale deployments

Service Name

AWS IoT FleetWise

Service Overview

AWS IoT FleetWise is a managed service that enables the collection, organization, and transformation of vehicle data in the cloud. The service helps improve vehicle quality, performance, and autonomy by collecting and standardizing data from vehicles using different protocols and data formats. It transforms low-level messages into human-readable values and keeps both structured and unstructured data synchronized in the cloud. FleetWise uses Edge Agent software running on vehicles to collect data based on intelligent campaigns that specify what data to collect and when to transfer it to the cloud for analysis and machine learning applications.

Key Use Cases

- Train AI/ML models: Continuously improve machine learning models used for autonomous and advanced driver assistance systems by collecting data from production vehicles
- Enhance digital customer experience: Use data from infotainment systems to make in-vehicle audiovisual content and in-app insights more relevant
- Maintain vehicle fleet health: Use insights from fleet data to monitor EV battery health and charge levels, manage maintenance schedules, analyze fuel consumption, and identify potential maintenance issues
- Create and manage commands: Remotely send commands to vehicles from the cloud, such as locking doors or setting temperature within seconds
- Create and manage state templates: Provide mechanisms for vehicle owners to track the state of their vehicles through collected signal updates

Features

- **Vehicle Modeling Framework**: Define and model vehicles with their sensors and actuators in the cloud using signals, signal catalogs, vehicle models, and decoder manifests
- **Edge Agent Software**: Reference implementation for developing software to facilitate secure communication between vehicles and cloud
- **Intelligent Data Collection**: Campaign-based data collection with condition-based and time-based schemes to control what vehicle data to collect and when to transfer
- **Multi-Protocol Support**: Support for various network interfaces including Controller Area Network (CAN bus), On-board Diagnostic (OBD) II, and custom interfaces

- **Data Compression**: SNAPPY compression to save wireless bandwidth and reduce network traffic during data transmission
- **Multiple Data Destinations**: Send data to MQTT topics, Amazon S3, or Amazon Timestream with support for MQTT v3.1.1 and v5.0 specifications

Value Proposition

AWS IoT FleetWise provides unique value through intelligent data collection that sends only relevant data to the cloud, reducing transmission costs and bandwidth usage. The service offers standardized fleet-wide data analysis without requiring custom data collection systems, automatic synchronization of both structured sensor data and unstructured vision system data in the cloud. Key differentiators include edge data storage capabilities that forward data under optimal conditions (such as Wi-Fi connectivity), built-in data transformation from binary to human-readable formats, and seamless integration with AWS analytics and machine learning services for advanced vehicle insights and autonomous driving improvements.

Service Integration

AWS IoT FleetWise integrates deeply with core AWS services to provide comprehensive vehicle data management. Primary integrations include AWS IoT Core for device provisioning and secure MQTT communication, Amazon Timestream for real-time time-series data storage and analytics, and Amazon S3 for cost-effective long-term data storage with Apache Parquet or JSON formats:

- Additional integrations support the broader ecosystem: Amazon SageMaker for machine learning model training, AWS Glue for data processing and cataloging, Amazon Athena for data querying, and AWS Step Functions for workflow orchestration
- The service also leverages AWS Key Management Service (KMS) for encryption and works with Amazon Macie for sensitive data protection

Onboarding and Offboarding

Customers get started with AWS IoT FleetWise by first setting up an AWS account and creating administrative users. The onboarding process involves deploying Edge Agent software using provided CloudFormation templates, creating vehicle models to standardize vehicle formats, establishing signal catalogs to define vehicle data structures, and setting up decoder manifests to transform binary data into readable values:

- Customers then create virtual vehicle representations, associate them with fleets, and configure data collection campaigns
- The service provides comprehensive tutorials, SDK examples in multiple languages (Java, Kotlin), and integration guides for connecting with existing AWS IoT Core infrastructure and downstream analytics services

Getting Started Guide: https://docs.aws.amazon.com/iot-fleetwise/latest/developerguide/fleetwise-getting-started.html

Data Removal

AWS IoT FleetWise enables customers to remove their data through standard AWS data deletion processes. Customers can delete campaigns, vehicles, decoder manifests, model manifests, fleets, and signal catalogs using provided API operations or console interfaces. Data stored in integrated services like Amazon S3 and Amazon Timestream can be removed using those services' respective data deletion capabilities. The service provides CloudFormation stack deletion options to remove all created resources and avoid further charges during offboarding.

Backup/Restore and Disaster Recovery

AWS IoT FleetWise operates as a managed service with built-in resilience across multiple availability zones. The service does not provide traditional backup capabilities but relies on underlying AWS infrastructure durability. Data collected and stored in Amazon S3 and Amazon Timestream inherits those services' backup and disaster recovery capabilities. Customers can implement cross-region data replication using S3 cross-region replication and Timestream's multi-region features for additional data protection and disaster recovery scenarios.

Backup Capabilities

AWS IoT FleetWise does not have native backup capabilities as it is a data collection and transformation service. However, data collected through campaigns is automatically stored in durable AWS services like Amazon S3 and Amazon Timestream, which provide their own backup and versioning capabilities. Customers can configure S3 lifecycle policies and cross-region replication for backup strategies, while Timestream provides built-in durability and retention policies.

Disaster Recovery

AWS IoT FleetWise supports disaster recovery through its multi-availability zone architecture and integration with resilient AWS services. The service does not directly integrate with AWS Elastic Disaster Recovery but benefits from AWS infrastructure resilience. Customers can implement disaster recovery strategies using connected mobility reference architectures that include cross-region replication of IoT device registries, certificates, and policies, along with data replication in S3 and Timestream across regions.

Recovery Objectives

AWS IoT FleetWise helps customers meet recovery objectives through its distributed architecture and integration with durable AWS services. The service provides near real-time data collection and transfer, enabling low RTO for data availability. RPO objectives are

supported through continuous data collection campaigns and immediate transfer to persistent storage in S3 and Timestream. Edge data partitioning capabilities provide additional data resilience by temporarily storing data locally before cloud transfer.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/general/latest/gr/iotfleetwise.html

FAQ: https://aws.amazon.com/iot-fleetwise/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/iot-fleetwise/latest/developerguide/what-is-iotfleetwise.html

User Guide: https://docs.aws.amazon.com/iot-fleetwise/latest/developerguide/

API Reference: https://docs.aws.amazon.com/iot-fleetwise/latest/APIReference/

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/iot-fleetwise/pricing/

Cost Optimization

AWS IoT FleetWise offers several unique cost optimization features including intelligent data collection that transmits only relevant data based on configurable conditions and time periods, reducing bandwidth costs. SNAPPY compression reduces data transfer volumes by up to 70%:

- Edge data partitioning allows temporary local storage and batch transmission during optimal network conditions like Wi-Fi connectivity
- Customers pay only for active vehicles and messages sent, with no minimum fees
- Data destination flexibility allows choosing cost-effective storage in S3 for archival or Timestream for real-time analytics based on use case requirements

Savings Considerations

Main cost savings come from reduced data transmission through intelligent filtering and compression, eliminating unnecessary raw data transfer. Edge processing and selective data collection minimize cloud storage and processing costs:

- Batch data transfer optimization reduces cellular data charges
- The pay-per-active-vehicle model avoids costs for inactive fleet members

- Integration with cost-effective AWS services like S3 for long-term storage provides significant savings compared to proprietary vehicle data platforms
- Automated data lifecycle management and the ability to configure retention policies help optimize storage costs over time while maintaining compliance with data retention requirements

Service Name

AWS IoT Greengrass

Service Overview

AWS IoT Greengrass is an open-source IoT edge runtime and cloud service that extends AWS IoT Core capabilities to edge devices. It enables devices to act locally on generated data, run machine learning models, filter and aggregate device data, and communicate securely with other devices and AWS IoT Core. The service provides local compute, messaging, data caching, and ML inference capabilities for connected devices, allowing them to operate even with intermittent cloud connectivity. AWS IoT Greengrass V2 introduces modular applications called components that can be deployed using Lambda functions, Docker containers, native OS processes, or custom runtimes.

Key Use Cases

- Industrial IoT and Edge Computing: Deploy edge applications for predictive maintenance, factory automation, and equipment monitoring in manufacturing environments
- Smart Cities and Infrastructure: Enable autonomous operation of traffic systems, smart lighting, and environmental monitoring with local data processing
- Connected Vehicles and Transportation: Support autonomous vehicle operations, fleet management, and real-time vehicle diagnostics with offline capability
- Agriculture and Environmental Monitoring: Implement precision agriculture solutions, greenhouse automation, and environmental sensing with intermittent connectivity
- Healthcare and Remote Monitoring: Deploy patient monitoring systems, medical device management, and healthcare analytics at the edge with secure local processing

Features

- **Local Processing**: Run AWS Lambda functions locally, support containers, AWS IoT Device Shadows, local messaging, and local resource access
- **Machine Learning Inference**: Deploy and run ML models at the edge with optimized performance using Amazon SageMaker and Neo compiler integration
- **Stream Manager**: Manage and process data streams from multiple sources with local buffering and cloud synchronization
- **Over-the-Air Updates**: Deploy software updates, configuration changes, and new components remotely to edge devices
- **Security Integration**: Hardware Security Module integration, secrets management, and end-to-end encryption for secure edge operations

- **Client Device Support**: Act as secure MQTT endpoint for local devices without direct cloud access, with local messaging and shadow synchronization
- **Component System**: Modular software architecture using pre-built components for extending device functionality and building applications
- **Offline Operation**: Continue local processing, messaging, and data caching when internet connectivity is intermittent or unavailable

Value Proposition

AWS IoT Greengrass provides unique edge computing capabilities that reduce latency, bandwidth costs, and dependency on cloud connectivity while maintaining security and scalability. It offers seamless integration with AWS IoT Core and other AWS services, enabling hybrid cloud-edge architectures:

- The service supports both lightweight (nucleus lite) and full-featured (nucleus) runtimes, accommodating resource-constrained to powerful edge devices
- Key advantages include local ML inference with optimized performance, offline operation capabilities, centralized fleet management, and extensive hardware ecosystem support
- Organizations benefit from reduced operational costs through local data processing, improved responsiveness with sub-millisecond latency for critical applications, and enhanced reliability with autonomous edge operation during connectivity issues

Service Integration

AWS IoT Greengrass integrates deeply with AWS IoT Core for device management, messaging, and shadow synchronization, serving as the primary cloud connection point. It connects with AWS Lambda for serverless edge computing, Amazon S3 for component artifacts and data storage, and AWS Secrets Manager for secure credential management:

- The service integrates with Amazon SageMaker and AWS IoT Analytics for ML model deployment and data analysis
- CloudWatch provides logging and monitoring capabilities, while AWS IAM handles authentication and authorization
- Additional integrations include Amazon Kinesis for data streaming, AWS Step Functions for workflow orchestration, Amazon DynamoDB for edge data storage, and AWS Systems Manager for configuration management
- The service also works with AWS IoT Device Management for fleet provisioning and AWS CloudFormation for infrastructure deployment

Onboarding and Offboarding

Getting started with AWS IoT Greengrass requires an AWS account, creating an AWS IoT thing for the core device, and installing the Greengrass Core software on a Linux or Windows device with minimum 512MB RAM and 1GB storage. The automated device setup

script simplifies initial configuration by installing dependencies, downloading certificates, and creating the Greengrass group:

- Alternatively, customers can follow the step-by-step tutorial covering environment setup, core device configuration, component development, and deployment testing
- Prerequisites include Python 3.6+, Java Runtime Environment, and basic familiarity with IoT concepts
- The service provides comprehensive documentation, video tutorials, and hands-on labs to accelerate adoption
- Fleet provisioning capabilities enable large-scale deployments with zero-touch device onboarding

Getting Started Guide:
https://docs.aws.amazon.com/greengrass/v2/developerguide/getting-started.html

Data Removal

When offboarding from AWS IoT Greengrass, customers must delete components and deployments from the cloud service, uninstall the Greengrass Core software from edge devices, and remove associated AWS IoT things, certificates, and policies. Local data stored on edge devices should be manually deleted, including component artifacts, logs, and cached data in the Greengrass work directories. Cloud-side data removal involves deleting component versions, deployment configurations, and associated CloudWatch logs. Customers should revoke device certificates and remove IAM roles and policies created for Greengrass operations to ensure complete data cleanup.

Backup/Restore and Disaster Recovery

AWS IoT Greengrass provides built-in resilience through local data caching, persistent sessions with AWS IoT Core, and automatic reconnection capabilities. For backup and recovery, the service supports local storage destinations for critical data, component artifact backup to S3, and configuration backup through Infrastructure as Code. Disaster recovery solutions include multi-region device registry replication, automated failover using Amazon Route 53, and cross-region component deployment capabilities. The AWS Disaster Recovery for AWS IoT solution provides automated replication of device shadows, certificates, and policies to secondary regions for comprehensive DR coverage.

Backup Capabilities

AWS IoT Greengrass includes native backup capabilities through local data storage, component artifact versioning in S3, and configuration backup via CloudFormation templates. The service supports local storage caching for resilience during connectivity issues and maintains persistent sessions for message delivery:

- While not directly integrated with AWS Backup, customers can backup device data using custom components and leverage S3 versioning for component artifacts

- Stream Manager provides local buffering and data persistence, while local shadows maintain device state information for recovery purposes

Disaster Recovery

AWS IoT Greengrass supports disaster recovery through multi-region deployment capabilities, automated device registry replication, and cross-region failover mechanisms. The service works with Amazon Route 53 health checks for automatic endpoint switching and supports AWS Application Recovery Controller for coordinated disaster recovery orchestration. While not directly integrated with AWS Elastic Disaster Recovery, customers can implement comprehensive DR strategies using AWS Step Functions, Lambda, and DynamoDB Global Tables for device registry replication across regions with automated failover capabilities.

Recovery Objectives

AWS IoT Greengrass helps customers achieve low RPO and RTO objectives through local data persistence, automatic reconnection capabilities, and multi-region deployment support. The service's offline operation capabilities minimize data loss during outages, while persistent MQTT sessions ensure message delivery upon reconnection. Stream Manager's local buffering reduces RPO to near-zero for critical data, and automated deployment mechanisms enable rapid service restoration. Cross-region replication and automated failover capabilities support RTO objectives of hours to minutes depending on the DR strategy implemented.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/general/latest/gr/greengrassv2.html

FAQ: https://aws.amazon.com/greengrass/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/greengrass/v2/developerguide/what-is-iot-greengrass.html

User Guide: https://docs.aws.amazon.com/greengrass/v2/developerguide/getting-started.html

API Reference: https://docs.aws.amazon.com/greengrass/v2/APIReference/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/greengrass/pricing/

Cost Optimization

AWS IoT Greengrass offers unique cost optimization through edge data processing that reduces cloud computing and data transfer costs. Local ML inference eliminates cloud API calls for real-time predictions, while Stream Manager's data aggregation and compression minimize bandwidth usage:

- The service's offline operation capabilities reduce dependency on continuous cloud connectivity, lowering data transfer costs
- Component-based architecture enables selective deployment of only required functionality, optimizing resource usage
- Free tier includes three Greengrass cores for one year, and pay-per-active-core pricing model aligns costs with actual usage
- Local data filtering and preprocessing reduce cloud storage and processing costs by sending only relevant data to AWS services

Savings Considerations

Primary cost savings with AWS IoT Greengrass come from reduced data transfer costs through local processing and selective data transmission to the cloud. Edge ML inference eliminates expensive cloud API calls for real-time predictions, while local data aggregation reduces cloud storage and processing costs:

- The service minimizes bandwidth requirements through data compression and filtering at the edge, particularly beneficial for remote deployments with expensive connectivity
- Operational savings result from centralized fleet management, automated updates, and reduced need for on-site technical support
- Free tier offering and usage-based pricing model allow cost-effective scaling
- Local processing capabilities reduce dependence on cloud resources for time-sensitive operations, optimizing overall infrastructure costs while maintaining performance requirements

Service Name

AWS IoT SiteWise

Service Overview

AWS IoT SiteWise is a managed service that enables industrial enterprises to collect, store, organize, and monitor data from industrial equipment at scale to make better, data-driven decisions. It provides a resource modeling framework to create representations of industrial devices, processes, and facilities called asset models. The service ingests data through multiple methods including direct OPC UA connections, edge processing, MQTT messages, and APIs, then processes and transforms this raw industrial data into actionable insights. It offers three storage tiers for equipment data and includes capabilities for real-time monitoring, predictive analytics, and visualization through managed web applications.

Key Use Cases

- Manufacturing: Monitor operations across facilities, compute industrial performance metrics like OEE and MTBF, prevent costly equipment issues and reduce production gaps
- Food and beverage: Optimize production processes, ensure quality control, monitor equipment performance and maintain compliance with safety standards
- Energy and utilities: Monitor power generation equipment, optimize asset performance, predict equipment failures and manage distributed energy resources

Features

- **Data Ingestion**: Collect data through OPC UA, MQTT, edge gateways, AWS IoT Core rules, and API methods with support for buffered ingestion
- **Asset Modeling**: Create representations of industrial devices, processes, and facilities with asset models, hierarchies, and property definitions
- **Edge Processing**: Process, store, and analyze data locally at the edge using SiteWise Edge with data collection and processing packs
- **Time Series Storage**: Three-tier storage system (hot, warm, cold) with intelligent data lifecycle management for cost optimization
- **Analytics and Monitoring**: SQL queries, real-time dashboards, alarms through CloudWatch, and anomaly detection for predictive maintenance
- **Visualization**: No-code web applications through SiteWise Monitor and Grafana plugin for operational data visualization
- **AI Assistant**: Generative AI-powered industrial assistant for insights, troubleshooting, and operational guidance

Value Proposition

AWS IoT SiteWise provides a complete industrial IoT solution that eliminates the complexity of collecting and contextualizing industrial data at scale. Unlike building custom solutions, it offers managed infrastructure with automatic scaling, built-in security, and integration with the broader AWS ecosystem:

- The service reduces time to value through pre-built asset modeling frameworks, automated data processing, and ready-to-use visualization tools
- Its edge processing capabilities enable hybrid architectures for low-latency requirements while maintaining cloud connectivity
- The three-tier storage system optimizes costs automatically, and the AI-powered assistant accelerates troubleshooting and operational insights, making it ideal for enterprises seeking rapid industrial digital transformation

Service Integration

AWS IoT SiteWise integrates deeply with AWS IoT Core for device connectivity and MQTT messaging, Amazon S3 for data lake storage and bulk operations, and Amazon CloudWatch for alarms and monitoring. It connects with AWS IoT Greengrass for edge computing capabilities and AWS IoT Events for event-driven data processing:

- The service leverages Amazon Lookout for Equipment for predictive maintenance analytics and AWS IoT TwinMaker for digital twin synchronization
- Additional integrations include AWS Lambda for custom processing, Amazon DynamoDB for operational data storage, Amazon Athena for SQL analytics, Amazon QuickSight for business intelligence, and Grafana for advanced visualization and monitoring dashboards

Onboarding and Offboarding

Customers can begin with AWS IoT SiteWise through a quick start demo that creates sample asset models, assets, and a SiteWise Monitor portal with generated data. The service offers multiple tutorials covering data ingestion from IoT things, SiteWise Monitor visualization, and integration with other AWS services:

- Getting started requires an AWS account and selecting a supported region
- Customers can choose from various data ingestion methods including direct OPC UA connections, edge gateways, or MQTT through IoT Core
- The console-based asset modeling allows users to define equipment representations, while SiteWise Monitor provides immediate visualization capabilities for operational teams

Getting Started Guide: https://docs.aws.amazon.com/iot-sitewise/latest/userguide/getting-started.html

Data Removal

Offboarding involves deleting AWS IoT SiteWise resources including assets, asset models, gateways, and portals through the console or APIs. Historical asset property data can be exported before deletion using query APIs or bulk export to Amazon S3. The quick start demo automatically deletes itself after a specified period, while production resources require manual cleanup. Resource definitions can be backed up using Describe APIs before deletion to enable recreation if needed.

Backup/Restore and Disaster Recovery

AWS IoT SiteWise leverages highly available AWS services like Amazon S3 and operates across multiple availability zones for inherent resilience. Data backup is achieved through property value notifications to IoT Core with rules routing to S3 and DynamoDB. The service provides APIs to retrieve and backup historical asset property data and resource definitions for disaster recovery scenarios.

Backup Capabilities

AWS IoT SiteWise does not natively integrate with AWS Backup service. However, it provides backup capabilities through property notifications that publish MQTT messages to AWS IoT Core, which can be routed to storage services like Amazon S3 and DynamoDB using IoT Core rules for data persistence and backup purposes.

Disaster Recovery

AWS IoT SiteWise does not directly integrate with AWS Elastic Disaster Recovery. The service achieves disaster recovery through its multi-AZ architecture and integration with durable AWS services. Customers can implement disaster recovery by replicating IoT Core device registries, certificates, and policies to secondary regions using custom solutions with DynamoDB Global Tables and Lambda functions.

Recovery Objectives

AWS IoT SiteWise helps meet RPO and RTO objectives through its multi-AZ architecture providing high availability within regions. Data replication to other AWS services enables backup with RPO measured in minutes. For cross-region disaster recovery, customers can achieve RTO of hours through automated replication of device registries and asset configurations, combined with data export capabilities to restore historical information.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/iot-sitewise/latest/userguide/endpoints-and-quotas.html

FAQ: https://aws.amazon.com/iot-sitewise/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/iot-sitewise/latest/userguide/what-is-sitewise.html

User Guide: https://docs.aws.amazon.com/iot-sitewise/latest/userguide/

API Reference: https://docs.aws.amazon.com/iot-sitewise/latest/APIReference/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/iot-sitewise/pricing/

Cost Optimization

AWS IoT SiteWise offers intelligent three-tier storage (hot, warm, cold) that automatically moves historical data to cost-optimized tiers based on user-defined policies, reducing storage costs significantly. Buffered ingestion allows defining data velocities for different use cases, optimizing messaging costs:

- Edge processing with SiteWise Edge reduces cloud compute and data transfer costs by processing data locally
- The service offers different pricing for near real-time versus buffered data ingestion, allowing cost optimization based on latency requirements
- Data export to S3 provides additional cost-effective storage options for long-term archival and analytics workloads

Savings Considerations

Primary cost savings come from the automated storage tiering system that keeps recent data in hot storage while moving historical data to lower-cost warm and cold tiers. Edge processing reduces cloud data transfer and compute costs by handling processing locally:

- Buffered ingestion provides lower-cost data ingestion compared to real-time streaming for non-critical use cases
- The managed service eliminates infrastructure overhead and maintenance costs compared to self-managed solutions
- Integration with other AWS services enables economies of scale and avoids data egress charges
- Predictive maintenance capabilities enabled by the service can prevent costly equipment failures and reduce operational expenses across industrial operations

Service Name

AWS Key Management Service

Service Overview

AWS Key Management Service (KMS) is a fully managed cryptographic service that enables customers to create and control encryption keys used to protect their data. The service operates using FIPS 140-3 Security Level 3 validated hardware security modules (HSMs) and provides centralized key management across AWS services and applications. KMS offers traditional key management services integrated with AWS, handling infrastructure management including availability, physical security, logical access control, and maintenance. Keys are protected within HSMs and never leave the service unencrypted, ensuring maximum security for data protection both at rest and in transit.

Key Use Cases

- Data encryption at rest: Protecting sensitive data stored in AWS services like S3, EBS, RDS, and DynamoDB using customer-managed or AWS-managed keys
- Client-side encryption: Using envelope encryption with AWS Encryption SDK to protect data before sending it to AWS services
- Digital signing and verification: Creating asymmetric keys for signing documents, APIs, and code, enabling authentication and integrity verification
- Compliance and regulatory requirements: Meeting standards like PCI DSS, HIPAA, and GDPR through centralized key management and comprehensive audit logging
- Multi-region data protection: Using multi-region keys to encrypt data across multiple AWS regions while maintaining consistent access controls

Features

- **Symmetric Encryption Keys**: 256-bit AES keys for encrypting and decrypting data up to 4 KB directly or for envelope encryption of larger datasets
- **Asymmetric Keys**: RSA, ECC, ML-DSA, and SM2 key pairs for encryption/decryption, digital signing, and key agreement operations
- **HMAC Keys**: Symmetric keys for generating and verifying hash-based message authentication codes for data integrity
- **Hardware Security Modules**: FIPS 140-3 Security Level 3 validated HSMs protecting keys with tamper-evident and tamper-resistant hardware
- **Automatic Key Rotation**: Annual automatic rotation for AWS-managed keys and configurable rotation for customer-managed keys
- **Multi-Region Keys**: Keys that can be replicated across multiple AWS regions for global applications and disaster recovery

- **Custom Key Stores**: Integration with AWS CloudHSM clusters or external key managers for additional control and compliance
- **Grants and Key Policies**: Fine-grained access controls using IAM policies, key policies, and temporary grants for delegation
- **CloudTrail Integration**: Comprehensive logging of all key usage and management operations for audit and compliance
- **Post-Quantum Cryptography**: Support for ML-DSA quantum-resistant digital signature algorithms for future-proof security

Value Proposition

AWS KMS provides unparalleled security through FIPS 140-3 validated HSMs while eliminating the complexity and cost of managing cryptographic infrastructure. Unlike self-managed solutions, KMS offers seamless integration with over 100 AWS services, automatic scaling, and built-in high availability across multiple regions:

- The service reduces operational overhead by handling key lifecycle management, rotation, and backup automatically
- KMS provides superior cost efficiency with pay-per-use pricing, no upfront hardware costs, and a generous free tier
- The centralized audit trail through CloudTrail integration simplifies compliance reporting, while envelope encryption enables protection of unlimited data volumes without performance impact

Service Integration

AWS KMS integrates natively with over 100 AWS services for seamless encryption. Core integrations include Amazon S3 for object-level encryption, Amazon EBS for volume encryption, Amazon RDS for database encryption, and Amazon DynamoDB for table-level protection:

- Container services like Amazon ECS and EKS use KMS for secrets management, while Amazon CloudTrail, AWS Config, and AWS Systems Manager Parameter Store rely on KMS for log and configuration encryption
- Analytics services including Amazon Redshift, Amazon EMR, and Amazon Athena leverage KMS for data warehouse and big data encryption
- The AWS Encryption SDK enables client-side encryption across applications, and services like AWS Lambda, Amazon API Gateway, and AWS Step Functions use KMS for serverless application security

Onboarding and Offboarding

Getting started with AWS KMS requires minimal setup through the AWS Management Console, CLI, or SDKs. Customers begin by creating their first KMS key, selecting between symmetric encryption, asymmetric signing, or HMAC keys based on use case:

- The service provides comprehensive SDK examples in Java, Python, and PHP for common operations like key creation, encryption, and policy management
- AWS offers hands-on labs, documentation, and getting started guides to accelerate adoption
- Integration with existing AWS services is typically achieved by specifying a KMS key ARN in service configurations
- The AWS KMS console provides intuitive key management, policy configuration, and monitoring capabilities for both technical and non-technical users

Getting Started Guide:
https://docs.aws.amazon.com/kms/latest/developerguide/example_kms_Scenario_Basics_section.html

Data Removal

AWS KMS supports secure key deletion with a mandatory 7-30 day waiting period to prevent accidental data loss. During this period, keys cannot be used for cryptographic operations but can be recovered if needed. For immediate cessation of key usage, customers can disable keys rather than delete them. When deleting imported key material, the key becomes immediately unusable but can be restored by re-importing the same material. Multi-region keys require deletion from the primary region to remove all replicas. Customer data encrypted with KMS keys remains accessible until the key is permanently deleted.

Backup/Restore and Disaster Recovery

AWS KMS provides built-in disaster recovery through automatic replication of keys and policies across multiple Availability Zones within a region. Multi-region keys enable cross-region replication for global disaster recovery scenarios. The service maintains durable storage of key metadata and policies with 99.999999999% durability. Key rotation creates new key versions while maintaining access to previous versions for decrypting older data. AWS Backup integrates with KMS to encrypt backup data, and customers can use customer-managed keys with AWS Backup's logically air-gapped vaults for additional protection.

Backup Capabilities

AWS KMS keys are automatically backed up across multiple Availability Zones with high durability storage. The service maintains historical key versions through automatic and on-demand rotation, enabling recovery of data encrypted with previous key versions:

- Multi-region keys provide cross-region backup capabilities for disaster recovery
- AWS KMS integrates with AWS Backup service, allowing customers to use customer-managed keys to encrypt backup vaults and recovery points

- Key metadata, policies, and grants are automatically replicated and maintained by the service infrastructure

Disaster Recovery

AWS KMS supports comprehensive disaster recovery through multi-region key replication and automatic failover capabilities. The service maintains keys across multiple Availability Zones within regions and supports cross-region replication via multi-region keys:

- Integration with AWS Application Recovery Controller enables automated disaster recovery orchestration
- KMS provides consistent global endpoints and automatic scaling during regional failures
- The service's distributed architecture ensures continued operation even during significant infrastructure events, with sub-second recovery times for most operations

Recovery Objectives

AWS KMS enables customers to achieve aggressive RPO and RTO objectives through multi-region keys and real-time replication. Multi-region keys provide near-zero RPO by maintaining synchronized copies across regions. RTO targets of minutes can be achieved through automatic failover mechanisms and globally distributed endpoints. The service's high availability architecture with 99.9% SLA ensures minimal downtime. Integration with AWS Application Recovery Controller automates recovery processes, while envelope encryption patterns enable rapid recovery of large datasets without re-encryption.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/general/latest/gr/kms.html

FAQ: https://aws.amazon.com/kms/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/kms/latest/developerguide/

User Guide: https://docs.aws.amazon.com/kms/latest/developerguide/overview.html

API Reference: https://docs.aws.amazon.com/kms/latest/APIReference/

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/kms/pricing/

Cost Optimization

AWS KMS offers several unique cost optimization opportunities: utilize the 20,000 free requests per month across all regions; implement data key caching with the AWS Encryption SDK to reduce API calls by up to 99%; use Amazon S3 bucket keys to reduce KMS requests for S3 encryption by sharing data keys across objects; leverage AWS-managed keys when customer control isn't required as they incur no monthly charges; optimize by using a single KMS key across multiple AWS services rather than creating separate keys; consider geographic placement as cross-region key usage incurs additional charges; implement request batching and connection pooling to minimize API overhead.

Savings Considerations

Primary cost savings strategies include leveraging AWS-managed keys for standard use cases eliminating $1/month per key charges; implementing AWS Encryption SDK with data key caching reducing request volumes by 90%+; using S3 bucket keys sharing encryption keys across objects reducing per-object KMS calls; consolidating multiple applications under fewer KMS keys while maintaining security boundaries; utilizing the generous free tier of 20,000 requests monthly; optimizing request patterns through batching and caching; avoiding unnecessary key rotations beyond security requirements; properly sizing custom key store usage; and monitoring CloudTrail costs associated with KMS API logging to optimize audit trail expenses.

Service Name

AWS Lake Formation

Service Overview

AWS Lake Formation is a fully managed service that simplifies building, securing, and managing data lakes on AWS. It provides centralized governance, security, and sharing capabilities for analytics and machine learning workloads by allowing fine-grained access control for data stored in Amazon S3 and metadata in AWS Glue Data Catalog. Lake Formation uses its own permissions model that complements IAM permissions, enabling secure access to data lakes and external data sources. The service automates complex data lake setup tasks including data ingestion, cataloging, transformation, and security management while providing comprehensive audit logging and compliance capabilities.

Key Use Cases

- Data Lake Governance: Centrally govern and secure enterprise data lakes with fine-grained access control at database, table, column, row, and cell levels
- Cross-Account Data Sharing: Securely share data internally and externally across AWS accounts with tag-based and named resource access controls
- Analytics and Machine Learning: Enable secure data access for analytics services like Amazon Athena, Amazon Redshift, Amazon EMR, and Amazon SageMaker
- Compliance and Audit: Implement comprehensive data access monitoring, audit logging, and regulatory compliance for sensitive data
- Data Cataloging: Automatically discover, catalog, and label data using AWS Glue crawlers with metadata management capabilities

Features

- **Fine-grained Access Control**: Database, table, column, row, and cell-level permissions with tag-based and attribute-based access control
- **Cross-Account Data Sharing**: Secure data sharing across AWS accounts and external organizations with centralized governance
- **Data Catalog Management**: Centralized metadata management using AWS Glue Data Catalog with automatic schema discovery
- **Audit and Compliance**: Comprehensive data access logging, monitoring, and audit capabilities for regulatory compliance
- **Blueprint Workflows**: Pre-built ETL workflows for database snapshots, incremental updates, and log file processing
- **Hybrid Access Mode**: Selective onboarding of Lake Formation permissions alongside existing IAM controls

Value Proposition

AWS Lake Formation eliminates the complexity of building and managing data lakes by providing a fully managed service with centralized governance. Unlike traditional approaches requiring manual setup of permissions, cataloging, and security, Lake Formation automates these processes while providing enterprise-grade security features:

- Key advantages include fine-grained access control without complex IAM policies, seamless integration with AWS analytics services, simplified cross-account data sharing, and comprehensive audit capabilities
- The service reduces operational overhead, accelerates time-to-value for data initiatives, and ensures consistent security policies across the organization while maintaining compatibility with existing AWS services and third-party tools

Service Integration

Lake Formation deeply integrates with AWS Glue and AWS Glue Data Catalog as its foundation for metadata management and ETL processing. It seamlessly works with analytics services including Amazon Athena, Amazon Redshift, Amazon EMR, and Amazon SageMaker for data access and processing:

- The service integrates with Amazon S3 for data storage and uses IAM for identity management
- Lake Formation also connects with Amazon Kinesis Data Firehose for data ingestion, AWS CloudTrail for audit logging, and supports external data sources through JDBC connections
- Additionally, it integrates with AWS Data Exchange for data sharing and supports third-party business intelligence tools and analytics platforms

Onboarding and Offboarding

Customers begin by setting up Lake Formation through the AWS Management Console or CloudFormation templates, starting with creating a data lake administrator and configuring the default permission model. The process involves registering Amazon S3 locations as data lake storage, creating databases in the Data Catalog, and setting up user permissions:

- Customers can use pre-built blueprints to create ETL workflows for data ingestion from JDBC sources or use AWS Glue crawlers for schema discovery
- Lake Formation provides tutorials for creating data lakes from CloudTrail logs or database sources, with step-by-step guidance for configuring security policies, cross-account sharing, and integrating with analytics services

Getting Started Guide: https://docs.aws.amazon.com/lake-formation/latest/dg/getting-started-setup.html

Data Removal

Lake Formation data removal involves deleting Data Catalog tables and databases, revoking all permissions granted to users and roles, and unregistering S3 locations from Lake Formation management. Customers must remove cross-account sharing configurations and delete any custom IAM roles created for Lake Formation workflows. Since Lake Formation manages metadata and permissions rather than storing actual data, the underlying S3 data remains intact unless explicitly deleted. Organizations should audit and revoke all Lake Formation permissions before decommissioning to ensure complete access removal.

Backup/Restore and Disaster Recovery

Lake Formation relies on the underlying AWS services for backup and disaster recovery capabilities. Data stored in Amazon S3 benefits from S3's durability and availability features, while metadata in AWS Glue Data Catalog is automatically replicated within the region. Lake Formation permissions and configurations can be backed up using AWS Config or exported via APIs for disaster recovery planning. The service supports cross-region data access but does not provide native backup functionality, requiring customers to implement backup strategies for S3 data and Data Catalog metadata.

Backup Capabilities

Lake Formation does not provide native backup capabilities and does not integrate with AWS Backup service. Backup strategies must be implemented at the underlying service level, primarily for S3 data using S3 Cross-Region Replication, versioning, or third-party solutions:

- Data Catalog metadata can be exported using AWS Glue APIs for backup purposes
- Customers should implement comprehensive backup strategies for both data and metadata components independently

Disaster Recovery

Lake Formation does not directly support AWS Elastic Disaster Recovery service integration. Disaster recovery must be planned at the infrastructure level, focusing on S3 data replication and Data Catalog reconstruction:

- The service supports cross-region data access, enabling disaster recovery architectures where data can be accessed from multiple regions
- Organizations must design disaster recovery strategies that encompass S3 replication, IAM role recreation, and Lake Formation permission reconfiguration in recovery regions

Recovery Objectives

Lake Formation can support customer RPO and RTO objectives through its integration with highly durable AWS services and cross-region capabilities. Since metadata is stored in AWS Glue Data Catalog with automatic replication, metadata RPO is typically measured in minutes. Data RPO depends on S3 replication strategies implemented by customers. RTO can be minimized through cross-region data access capabilities and automated infrastructure deployment using CloudFormation templates for permission and configuration recreation.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/general/latest/gr/lake-formation.html#limits_lake-formation

FAQ: https://aws.amazon.com/lake-formation/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/lake-formation/latest/dg/what-is-lake-formation.html

User Guide: https://docs.aws.amazon.com/lake-formation/latest/dg/getting-started-tutorials.html

API Reference: https://docs.aws.amazon.com/lake-formation/latest/dg/aws-lake-formation-api.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/lake-formation/pricing/

Cost Optimization

Lake Formation offers several unique cost optimization opportunities including no additional charges for basic permissions management and data governance features. The service provides significant cost savings through its Storage API pricing model where customers only pay for bytes scanned, encouraging the use of columnar formats like Parquet and ORC to reduce scanning costs:

- Governed Tables feature optimizes storage layout automatically to improve query performance and reduce costs
- The service eliminates the need for separate data governance tools and reduces operational costs through automation

- Additionally, Lake Formation's ability to enable secure data sharing can reduce data duplication across organizations

Savings Considerations

Primary cost savings include elimination of third-party data governance tool licensing fees and reduced operational overhead through automated data lake management. The service's integration with serverless AWS analytics services enables pay-per-query pricing models, avoiding fixed infrastructure costs:

- Storage optimization through automatic compaction and columnar format recommendations reduces both storage and compute costs
- Cross-account data sharing capabilities eliminate data duplication costs and reduce data transfer charges
- The unified security model reduces administrative overhead and compliance costs
- Organizations can achieve significant TCO reduction by consolidating data governance, reducing manual security management, and leveraging automated optimization features

Service Name

AWS Lambda

Service Overview

AWS Lambda is a serverless compute service that runs code without requiring users to provision or manage servers. It automatically scales execution environments to handle increased requests, manages underlying infrastructure including server maintenance and capacity provisioning, and charges based on actual usage. Lambda functions execute in response to events from AWS services or direct invocation, supporting multiple programming languages including Node.js, Python, Java, Go, and .NET. The service handles scaling, logging, and capacity management while users focus solely on their application code.

Key Use Cases

- File processing: Automatically process files when uploaded to Amazon S3 or other storage services
- Real-time stream processing: Process data streams from Kinesis, DynamoDB, and other event sources for analytics and monitoring
- Web and mobile backends: Create scalable API backends that automatically adjust to demand using API Gateway integration
- Scheduled tasks: Run automated operations on regular schedules using EventBridge for maintenance and data processing
- Database operations: Respond to database changes and automate data workflows with triggers from RDS and DynamoDB
- Long-running workflows: Build stateful, multi-step workflows using durable functions that can run for up to one year
- IoT backends: Handle IoT device requests and process sensor data at scale

Features

- **Automatic Scaling**: Automatically scales execution environments from zero to thousands based on demand with configurable concurrency controls
- **Multi-language Runtime Support**: Supports Node.js, Python, Java, Go, .NET, Ruby, and custom runtimes with managed runtime environments
- **Event-driven Architecture**: Integrates with 200+ AWS services as event sources including S3, DynamoDB, API Gateway, and EventBridge
- **Lambda Layers**: Share code libraries and dependencies across multiple functions to optimize code reuse and maintenance
- **Provisioned Concurrency**: Pre-initialize execution environments to eliminate cold starts for latency-sensitive applications

- **SnapStart**: Reduces cold start times to sub-second startup performance by creating snapshots during initialization
- **VPC Integration**: Access resources in Amazon VPC with network isolation and security controls
- **Container Image Support**: Deploy functions using container images up to 10GB for complex dependencies and custom runtimes
- **Response Streaming**: Stream responses progressively for improved performance in data-intensive applications
- **Code Signing**: Ensure code integrity and security compliance by verifying signed code before execution

Value Proposition

AWS Lambda offers compelling advantages over traditional server-based computing and other serverless alternatives. It eliminates server management overhead, provides automatic scaling from zero to enterprise scale, and offers true pay-per-use pricing with no charges for idle time:

- Lambda's extensive AWS service integrations enable rapid development of event-driven architectures
- The service provides sub-second cold start performance with SnapStart, enterprise-grade security with built-in isolation, and comprehensive monitoring through CloudWatch
- Organizations benefit from reduced operational complexity, faster time-to-market, and significant cost savings especially for variable workloads, making it ideal for modern application development

Service Integration

Lambda integrates natively with over 200 AWS services as event sources and destinations. Key integrations include Amazon API Gateway for web APIs, Amazon S3 for file processing, Amazon DynamoDB for database triggers, Amazon EventBridge for scheduled tasks, Amazon Kinesis for stream processing, Amazon SQS and SNS for messaging, Amazon RDS for database events, AWS Step Functions for workflow orchestration, Amazon CloudWatch for monitoring and logging, AWS IAM for security and permissions, Amazon VPC for network isolation, Amazon EFS for shared file systems, and AWS X-Ray for distributed tracing. These integrations enable comprehensive serverless architectures.

Onboarding and Offboarding

Getting started with Lambda is straightforward through the AWS Console, CLI, or Infrastructure as Code tools. Users begin by creating their first function using built-in templates or sample code, configuring triggers from AWS services, and testing using the console editor:

- The service provides extensive documentation, tutorials, and sample applications in multiple programming languages
- Deployment options include uploading ZIP files, container images, or using CI/CD pipelines
- AWS provides step-by-step guides, best practices documentation, and integration examples to help users build serverless applications efficiently
- No infrastructure provisioning or server management is required to begin development

Getting Started Guide: https://docs.aws.amazon.com/lambda/latest/dg/getting-started.html

Data Removal

Lambda function deletion removes all versions, aliases, and associated resources including execution roles and CloudWatch log groups. Functions can be deleted via console, CLI, or API with immediate effect. Source code and deployment packages are permanently removed from AWS storage. Event source mappings and triggers must be deleted separately. For complete offboarding, users should delete related resources like IAM roles, CloudWatch logs, and S3 deployment artifacts to ensure no residual data remains.

Backup/Restore and Disaster Recovery

Lambda provides built-in durability and availability across multiple Availability Zones within a region. Function code and configuration are automatically replicated for high availability. For disaster recovery, users can deploy functions across multiple regions using Infrastructure as Code. AWS Backup doesn't directly support Lambda functions, but deployment packages can be stored in S3 with cross-region replication. Version control and CI/CD pipelines serve as primary backup mechanisms for function code and configuration.

Backup Capabilities

Lambda functions don't require traditional backup as code is stored durably in the service. Function code, configurations, and versions are automatically replicated across Availability Zones:

- AWS Backup doesn't directly support Lambda functions
- Best practice involves storing deployment packages in S3 with versioning enabled and using version control systems for source code
- Infrastructure as Code templates provide configuration backup and enable consistent deployment across environments

Disaster Recovery

Lambda supports disaster recovery through multi-region deployment strategies. Functions can be deployed to multiple regions using Infrastructure as Code for active-active or active-passive configurations:

- AWS Elastic Disaster Recovery doesn't directly apply to Lambda as it's a managed serverless service
- Built-in durability within regions and cross-region deployment capabilities provide comprehensive disaster recovery
- Application Recovery Controller can orchestrate multi-region Lambda applications for automated failover scenarios

Recovery Objectives

Lambda enables aggressive RPO and RTO objectives through serverless architecture. With cross-region deployment, RPO can approach zero for stateless functions. RTO depends on DNS failover and traffic routing, typically achieving minutes to sub-minute recovery times. Built-in durability across Availability Zones provides sub-second recovery for regional failures. Provisioned concurrency and SnapStart minimize cold start impact during recovery. Global applications can achieve near-zero downtime with proper architecture using Route 53 health checks.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/lambda/latest/dg/gettingstarted-limits.html

FAQ: https://aws.amazon.com/lambda/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/lambda/latest/dg/welcome.html

User Guide: https://docs.aws.amazon.com/lambda/latest/dg/welcome.html

API Reference: https://docs.aws.amazon.com/lambda/latest/api/

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/lambda/pricing/

Cost Optimization

Lambda offers unique cost optimization through precise pay-per-use billing based on invocation count and execution duration. AWS Compute Optimizer provides ML-driven recommendations for optimal memory configurations to reduce costs and improve performance:

- Graviton processors can reduce costs by up to 20% with better performance
- SnapStart reduces cold start duration costs, and Provisioned Concurrency eliminates cold starts for predictable workloads
- Lambda Managed Instances provide cost savings for steady-state workloads through EC2 pricing options
- Efficient memory allocation, reduced package sizes, and optimized initialization code minimize execution costs
- Reserved capacity and Savings Plans can provide additional discounts for predictable usage patterns

Savings Considerations

Primary cost savings come from eliminating server provisioning and paying only for actual compute time used, not idle capacity. Short-duration, intermittent workloads see the most dramatic savings compared to traditional servers:

- Memory optimization can reduce costs by 10-50% through Compute Optimizer recommendations
- Efficient code and reduced cold starts minimize billable duration
- Container image reuse and layer optimization reduce deployment costs
- Multi-region strategies should consider data transfer costs
- Graviton processors offer 20% better price performance
- For high-volume predictable workloads, Lambda Managed Instances can provide significant savings over on-demand pricing while maintaining serverless benefits

Service Name

AWS License Manager

Service Overview

AWS License Manager is a service that helps organizations manage and track software licenses from various vendors, such as Microsoft, IBM, Oracle, and SAP, across multiple AWS regions and accounts. It offers consolidated visibility and comprehensive reporting for software license compliance at scale, reducing the risk of licensing overages and non-compliance. The service supports tracking of software that is licensed based on virtual cores, physical cores, sockets, or number of machines. License Manager allows administrators to set hard or soft limits on new and existing cloud deployments, helping prevent non-compliant server usage and provides built-in dashboards for ongoing visibility into license usage and assistance with vendor audits.

Key Use Cases

- Bring Your Own License (BYOL) management: Manage existing software licenses for use with cloud resources while maintaining compliance
- License compliance tracking: Monitor and enforce license usage across AWS and on-premises environments to prevent violations
- Automated license discovery: Automatically discover and track software installed on AWS resources including EC2 instances, RDS databases, and Elastic Beanstalk environments
- Vendor audit preparation: Provide comprehensive reporting and visibility to assist with software vendor audits and compliance verification
- Cross-account license management: Manage licenses across multiple AWS accounts and regions within an organization for centralized oversight

Features

- **Self-managed licenses**: Create and manage licensing rules based on enterprise agreements to determine how AWS processes license-consuming commands
- **License asset groups**: Provide organization-wide license management across multiple regions and accounts with centralized oversight and automated compliance monitoring
- **Automated discovery**: Automatically discover and track software installed on AWS resources, including uninstalled software tracking
- **Managed entitlements**: Enable ISVs to manage and distribute software licenses and data to end-users through centralized entitlement management
- **User-based subscriptions**: Associate licenses with specific users instead of instances for Microsoft Office and Visual Studio subscriptions

- **Linux subscriptions management**: Manage and track usage of Linux subscriptions including Red Hat and SUSE with conversion capabilities
- **License type conversions**: Convert instances from one license type to another and switch between different licensing models
- **Built-in dashboards**: Monitor license usage and compliance through comprehensive dashboards with real-time visibility

Value Proposition

AWS License Manager provides unique value by offering centralized license management across both cloud and on-premises environments with no additional charges. Key advantages include: automated compliance enforcement that prevents licensing violations through rule-based controls, significant cost savings through optimized BYOL opportunities and vCPU-based licensing cost reduction, comprehensive visibility with built-in dashboards for real-time license usage tracking, seamless integration with core AWS services like EC2, RDS, Systems Manager, and Organizations, and simplified vendor audit preparation with detailed reporting and compliance tracking. The service supports multiple licensing models including perpetual, floating, subscription, and usage-based licenses while using open industry standards for license representation and cryptographic verification.

Service Integration

AWS License Manager integrates deeply with core AWS services to provide comprehensive license management. Primary integrations include: Amazon EC2 for tracking instances, Dedicated Instances, Dedicated Hosts, and Spot Instances; Amazon RDS for Oracle and DB2 vCPU-based BYOL licenses; AWS Systems Manager for managed nodes and inventory collection; AWS Organizations for cross-account discovery and management; AWS IAM for access control and permissions; AWS Marketplace and AWS Data Exchange for license distribution; CloudFormation for infrastructure as code deployment; Service Quotas for limit management; AWS resource tagging for organization; and AWS X-Ray for monitoring. The service also works with AWS Managed Microsoft Active Directory for user subscription management and integrates with various launch mechanisms including EC2 launch templates and Service Catalog products.

Onboarding and Offboarding

Getting started with AWS License Manager involves opening the License Manager console and following prompts to configure permissions for License Manager and its supporting services. The initial setup process automatically configures necessary IAM roles and permissions:

- Customers can then create self-managed licenses by working with their compliance teams to review enterprise agreements and build appropriate licensing rules

- The service supports multiple onboarding paths including associating licenses with AMIs, using launch templates, CloudFormation templates, or Service Catalog products
- For organizations, administrators can enable trusted access with AWS Organizations to enable cross-account discovery and centralized management
- The setup process typically involves defining license terms as rules, configuring automated discovery for existing resources, and establishing monitoring through built-in dashboards

Getting Started Guide: https://docs.aws.amazon.com/license-manager/latest/userguide/getting-started.html

Data Removal

License Manager data removal involves deactivating and deleting self-managed licenses, disassociating license configurations from AMIs and resources, and removing organizational integrations. When offboarding, administrators should first deactivate licenses to stop tracking, then systematically disassociate licenses from all tracked resources including EC2 instances and RDS databases. For Organizations integration, administrators must disable trusted access using the Organizations CLI or API. License configurations are automatically disassociated from deregistered AMIs every 8 hours, and all license tracking data is treated as sensitive customer data without AWS access.

Backup/Restore and Disaster Recovery

AWS License Manager does not provide traditional backup and disaster recovery capabilities as it is a management service rather than a data storage service. The service manages license metadata and configurations which are maintained by AWS infrastructure. License configurations and rules are stored within the AWS service infrastructure with built-in redundancy. For disaster recovery, customers should document their license configurations and rules as Infrastructure as Code using CloudFormation templates to enable recreation in alternate regions if needed. The service relies on AWS's underlying infrastructure resilience for availability and data protection.

Backup Capabilities

AWS License Manager does not have traditional backup capabilities as it manages license metadata rather than customer data. The service does not integrate with AWS Backup since license configurations are management constructs maintained by the AWS service infrastructure:

- License tracking data and configurations are automatically maintained by the service with built-in redundancy across AWS availability zones
- Customers should maintain documentation of their license rules and configurations, potentially using Infrastructure as Code approaches for reproducibility across regions or accounts

Disaster Recovery

AWS License Manager does not directly support AWS Elastic Disaster Recovery as it is a management service focused on license tracking rather than application recovery. The service relies on AWS's underlying infrastructure for resilience and availability:

- For disaster recovery planning, customers should document license configurations and consider using CloudFormation templates to recreate license rules in alternate regions
- Cross-region license tracking is not supported, so customers must establish separate license configurations in each region where they operate

Recovery Objectives

AWS License Manager supports recovery objectives indirectly by maintaining license compliance during infrastructure recovery scenarios. The service ensures license rules and tracking continue functioning across AWS infrastructure failures through built-in service resilience. During recovery scenarios, license configurations remain available to prevent compliance violations when launching new instances. However, specific RPO/RTO objectives are not applicable since License Manager manages configuration metadata rather than critical business data. The service contributes to overall recovery by maintaining license compliance throughout infrastructure restoration processes.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/general/latest/gr/licensemanager.html

FAQ: https://aws.amazon.com/license-manager/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/license-manager/latest/userguide/license-manager.html

User Guide: https://docs.aws.amazon.com/license-manager/latest/userguide/

API Reference: https://docs.aws.amazon.com/license-manager/latest/APIReference/

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/license-manager/pricing/

Cost Optimization

AWS License Manager offers unique cost optimization through several mechanisms: BYOL cost savings by enabling customers to repurpose existing license inventory for cloud resources, potentially saving up to 50% on EC2 instance costs; vCPU-based licensing cost reduction through integration with Optimize CPU capability to track custom vCPU usage; automated license compliance to prevent costly licensing violations and penalties; license utilization optimization through tracking and reporting that identifies unused or underutilized licenses; Dedicated Hosts cost optimization for server-bound software licenses; and elimination of manual license tracking overhead. The service itself is provided at no additional charge, making license management cost-effective while enabling substantial savings through optimized license usage and compliance.

Savings Considerations

Primary cost savings from AWS License Manager include: significant BYOL savings through optimized license utilization and compliance, with potential reductions of up to 77% for Windows Server licenses and 45% for SQL Server licenses when migrating to AWS; prevention of costly licensing violations and audit penalties through automated compliance tracking; reduced administrative overhead by automating manual license management tasks; optimized resource utilization through right-sizing recommendations and license-to-resource matching; cost avoidance through proactive license limit enforcement preventing overages; and improved negotiating position with vendors through detailed usage analytics. Organizations can also achieve savings through centralized license management across multiple accounts and regions, reducing duplicate licensing costs and improving license pool utilization efficiency.

Service Name

AWS Network Firewall

Service Overview

AWS Network Firewall is a stateful, managed network firewall and intrusion detection and prevention service for Amazon Virtual Private Cloud (VPC). It filters traffic at the perimeter of VPCs, including traffic to and from internet gateways, NAT gateways, or over VPN or Direct Connect connections. The service uses the open source intrusion prevention system Suricata for stateful inspection and supports Suricata-compatible rules. Network Firewall enables organizations to monitor and protect VPC traffic through deep packet inspection, domain filtering, custom rule definitions, and protocol detection capabilities.

Key Use Cases

- VPC perimeter security: Filter network traffic at VPC boundaries to protect against external threats and control access to internet resources
- Compliance and regulatory requirements: Meet security standards through centralized firewall management and detailed traffic monitoring and logging
- Advanced threat protection: Block malicious URLs, command-and-control channels, and other threat vectors using intrusion prevention system capabilities
- Domain-based filtering: Control outbound traffic by allowing only known good domains or blocking known malicious domains
- Deep packet inspection: Perform comprehensive analysis of network traffic entering or leaving VPCs for security threats
- Multi-account security governance: Centrally manage firewall policies across multiple AWS accounts using AWS Firewall Manager integration

Features

- **Stateful and Stateless Inspection**: Supports both stateful rules (context-aware traffic flow inspection) and stateless rules (individual packet inspection) for comprehensive traffic filtering
- **Suricata-based IPS**: Uses open source Suricata intrusion prevention system for advanced threat detection and prevention with compatible rule support
- **TLS Inspection**: Advanced inspection capability that decrypts and inspects SSL/TLS encrypted traffic for enhanced security analysis
- **Active Threat Defense**: Managed rule groups providing real-time threat intelligence and protection against emerging threats
- **Domain and URL Filtering**: Web filtering capabilities to control access to specific domains and URLs, including automated domain list generation

- **Flow Management**: Flow capture and flush operations for network visibility, troubleshooting, and security policy enforcement
- **Central Management**: Integration with AWS Firewall Manager for centralized configuration and management across multiple AWS accounts
- **Multi-VPC Support**: Support for multiple VPC endpoints and associations for scalable network security architecture
- **Transit Gateway Integration**: Native integration with AWS Transit Gateway for simplified centralized network security deployment

Value Proposition

AWS Network Firewall provides a fully managed solution that eliminates the need to deploy, maintain, or patch network security infrastructure. Unlike self-managed solutions, it offers automatic scaling up to 100 Gbps throughput per Availability Zone with built-in high availability:

- The service integrates natively with AWS services like Transit Gateway, VPC, and Firewall Manager, providing seamless deployment and centralized management
- Cost-effective NAT Gateway discounts are available when used in service chains
- The managed Suricata-based engine provides enterprise-grade threat protection with continuously updated threat intelligence, while flexible rule management supports both custom and AWS Marketplace partner rule groups

Service Integration

Network Firewall integrates deeply with core AWS networking and security services. It works natively with Amazon VPC for subnet-level protection and route table management:

- Transit Gateway integration enables centralized network security across multi-VPC architectures without dedicated inspection VPCs
- AWS Firewall Manager provides centralized policy management across multiple accounts and Organizations
- The service integrates with NAT Gateways for cost-optimized outbound traffic inspection, and supports VPC endpoint associations for multi-VPC deployments
- Additional integrations include Route 53 Resolver DNS Firewall, AWS PrivateLink, CloudWatch for monitoring and logging, CloudTrail for audit trails, and AWS Certificate Manager for TLS inspection certificate management

Onboarding and Offboarding

Getting started requires having a VPC with internet gateway and designated firewall subnets in each Availability Zone. The process involves creating stateless and stateful rule groups that define traffic filtering behavior, creating a firewall policy that combines rule groups and specifies default actions, creating the firewall resource and associating it with

the VPC and subnets, and updating VPC route tables to direct traffic through firewall endpoints:

- The AWS Management Console provides step-by-step guidance, while CloudFormation templates and APIs enable automated deployments
- Prerequisites include proper IAM permissions and network architecture planning for firewall subnet placement and routing configurations

Getting Started Guide: https://docs.aws.amazon.com/network-firewall/latest/developerguide/getting-started.html

Data Removal

Offboarding involves reversing VPC route table changes to remove traffic routing through firewall endpoints, deleting VPC endpoint associations if multiple VPCs are protected, and systematically removing Network Firewall resources including the firewall, firewall policy, and rule groups. All firewall logs stored in CloudWatch Logs and S3 must be manually deleted based on organizational data retention policies. Shared resources managed through AWS Firewall Manager require coordination across accounts for proper cleanup and policy removal.

Backup/Restore and Disaster Recovery

AWS Network Firewall is a managed service where AWS handles infrastructure backup and recovery. Customer-configured resources like firewall policies, rule groups, and configurations can be backed up using Infrastructure as Code approaches with CloudFormation templates or Terraform. The service provides automatic multi-Availability Zone deployment for high availability, but does not integrate with traditional backup services like AWS Backup since it manages infrastructure-level resources rather than data that requires backup and restoration.

Backup Capabilities

Network Firewall does not have traditional backup capabilities or integration with AWS Backup since it manages infrastructure resources rather than data. Configuration backup is achieved through Infrastructure as Code practices, exporting resource configurations to CloudFormation templates, or using AWS Config for configuration history tracking. AWS manages the underlying infrastructure backup and recovery automatically, while customers are responsible for maintaining configuration templates and policies for disaster recovery scenarios.

Disaster Recovery

Network Firewall does not directly integrate with AWS Elastic Disaster Recovery as it is an infrastructure service rather than a data service. Disaster recovery is achieved through the service's built-in multi-Availability Zone deployment for high availability and Infrastructure as Code approaches for cross-region deployment. The managed nature of the service

means AWS handles underlying infrastructure recovery, while customers implement disaster recovery through automated deployment scripts, CloudFormation templates, and multi-region architecture patterns.

Recovery Objectives

Network Firewall supports aggressive RTO and RPO objectives through its multi-Availability Zone deployment architecture and Infrastructure as Code deployment capabilities. High availability across multiple AZs provides near-zero RTO for availability zone failures. Cross-region deployment using CloudFormation or Terraform can achieve RTO of minutes to hours depending on automation sophistication. RPO is effectively zero for configuration data when using Infrastructure as Code practices, as all configurations can be version-controlled and rapidly redeployed in disaster scenarios.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/network-firewall/latest/developerguide/quotas.html

FAQ: https://aws.amazon.com/network-firewall/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/network-firewall/latest/developerguide/what-is-aws-network-firewall.html

User Guide: https://docs.aws.amazon.com/network-firewall/latest/developerguide/

API Reference: https://docs.aws.amazon.com/network-firewall/latest/APIReference/

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/network-firewall/pricing/

Cost Optimization

Network Firewall offers several unique cost optimization opportunities. NAT Gateway hourly and data processing charges are waived when placed next to Network Firewall in the same region and Availability Zone service chain:

- Secondary VPC endpoints have reduced hourly rates compared to primary endpoints
- Advanced Inspection and Active Threat Defense features have separate pricing tiers, allowing customers to enable only required capabilities

- Flexible cost allocation through Transit Gateway attachments enables distributing inspection costs across AWS accounts based on actual usage rather than consolidating expenses in the firewall owner account

Savings Considerations

Primary cost savings come from the NAT Gateway discount when architecting service chains properly in the same AZ. Centralized deployment models reduce the need for multiple firewall instances across VPCs while maintaining security:

- Using Infrastructure as Code reduces operational overhead and management costs
- The managed service eliminates costs associated with self-managed firewall infrastructure, including patching, maintenance, and scaling operations
- Combining with VPC Lattice for non-critical HTTP/HTTPS traffic can significantly reduce Network Firewall processing costs while maintaining security
- Regional deployment optimization and right-sizing firewall endpoints based on actual throughput requirements prevent over-provisioning costs

Service Name

AWS Organizations

Service Overview

AWS Organizations is a centralized account management and governance service that helps businesses manage and govern their AWS resources as they scale. It allows users to create accounts, group accounts into organizational units (OUs), apply policies for governance, and simplify billing by using a single payment method for all accounts. Organizations integrates with other AWS services to define central configurations, security mechanisms, audit requirements, and resource sharing across accounts in your organization. The service provides consolidated billing, access control, policy management, and enables trusted access for AWS services to perform tasks on behalf of the organization.

Key Use Cases

- Automate AWS account creation and categorize workloads: Quickly launch new workloads by automating account creation and adding them to user-defined groups for instant security policy application, touchless infrastructure deployments, and auditing
- Define and enforce audit and compliance policies: Apply service control policies (SCPs) to ensure users perform only actions that meet security and compliance requirements, create central logs with AWS CloudTrail, and enforce standard configurations with AWS Config
- Provide tools and access for security teams: Create security groups with read-only access across all resources to identify and mitigate security concerns, manage Amazon GuardDuty for threat monitoring, and use IAM Access Analyzer to identify unintended resource access
- Share common resources across accounts: Share critical central resources like AWS Directory Service for Microsoft Active Directory as a central identity store, use AWS Service Catalog to share IT services, and share VPC subnets across the organization using AWS Resource Access Manager

Features

- **Multi-account Management**: Create, organize, and manage multiple AWS accounts programmatically through CLI, SDKs, or APIs with centralized provisioning of resources and permissions
- **Organizational Units (OUs)**: Group accounts into hierarchical organizational units to organize workflows, apply policies, and manage access controls at scale

- **Service Control Policies (SCPs)**: Define and enforce preventive guardrails that control access to AWS services, resources, and regions across accounts and OUs
- **Consolidated Billing**: Centralize billing and payment for all member accounts with volume discounts, combined usage benefits, and single bill management
- **Trusted Access Integration**: Enable AWS services to perform tasks on behalf of the organization through service-linked roles with predefined permissions
- **Resource Control Policies (RCPs)**: Centrally prevent unintended use of AWS resources with fine-grained access controls
- **Tag Policies**: Standardize resource tagging across accounts for classification, tracking, and attribute-based access control
- **Centralized Security and Monitoring**: Provide security teams with read-only access across accounts and centrally manage security tools like GuardDuty, Security Hub, and IAM Access Analyzer

Value Proposition

AWS Organizations provides unmatched centralized governance and management capabilities for multi-account AWS environments. Unlike managing accounts individually, Organizations offers automated account creation, hierarchical organization with OUs, and policy-driven governance through SCPs and RCPs:

- The service integrates seamlessly with 40+ AWS services for trusted access, enabling centralized security, compliance, and operational management
- Organizations delivers significant cost savings through consolidated billing, volume discounts, and shared Reserved Instances/Savings Plans across accounts
- The service scales effortlessly to manage up to 10,000 accounts while maintaining security isolation and granular access control
- With built-in compliance features, automated policy enforcement, and centralized audit capabilities through CloudTrail integration, Organizations provides enterprise-grade governance that manual account management cannot match

Service Integration

AWS Organizations integrates with over 40 AWS services through trusted access, enabling centralized management and governance. Key integrations include AWS IAM Identity Center for single sign-on across accounts, AWS Control Tower for automated multi-account setup with governance guardrails, and AWS CloudTrail for organization-wide activity logging:

- Security integrations encompass Amazon GuardDuty for threat detection, AWS Security Hub for centralized security findings, AWS Config for compliance monitoring, and IAM Access Analyzer for unintended access detection
- Cost management integrates with AWS Cost Explorer, AWS Billing and Cost Management, and AWS Compute Optimizer

- Resource sharing is enabled through AWS Resource Access Manager (RAM), AWS Service Catalog for IT service catalogs, and AWS License Manager for centralized licensing
- Additional integrations include AWS Backup for centralized backup policies, AWS Systems Manager for cross-account operations, and AWS CloudFormation StackSets for multi-account deployments

Onboarding and Offboarding

Getting started with AWS Organizations requires an AWS account with administrative access and MFA enabled on the root user. Create an organization by choosing between consolidated billing features only or all features (recommended):

- Begin by creating organizational units (OUs) to organize accounts hierarchically, then invite existing accounts or create new ones programmatically
- Configure service control policies (SCPs) to establish governance guardrails and enable trusted access for integrated AWS services like IAM Identity Center, CloudTrail, and Config
- Set up consolidated billing and cost allocation tags for financial management
- AWS provides a comprehensive tutorial for creating and configuring organizations, along with best practices for OU design, policy management, and security controls
- The AWS Management Console, CLI tools, SDKs, and APIs provide multiple access methods for managing the organization

Getting Started Guide:
https://docs.aws.amazon.com/organizations/latest/userguide/orgs_getting-started.html

Data Removal

AWS Organizations offboarding involves removing member accounts from the organization and optionally deleting the organization itself. Member accounts can be removed by the management account or leave voluntarily, retaining their resources and data. When removing accounts, ensure billing responsibilities are transferred and access permissions are updated. To delete an organization, first remove all member accounts, then delete organizational units and policies. Account closure follows AWS standard procedures with a 90-day retention period for closed accounts.

Backup/Restore and Disaster Recovery

AWS Organizations itself does not provide direct backup and disaster recovery capabilities as it's a management and governance service. However, it enables centralized backup and DR policies across member accounts through integration with AWS Backup for organization-wide backup policies and AWS Config for compliance monitoring. Organizations metadata and policies exist as managed service configurations that AWS maintains for high availability and durability across multiple availability zones.

Backup Capabilities

AWS Organizations does not have traditional backup capabilities as it's a governance service managing account relationships and policies. The service configuration and metadata are automatically maintained by AWS with high availability and durability. Organizations enables centralized backup governance through AWS Backup integration, allowing backup policies to be applied across accounts and OUs, but does not directly integrate with AWS Backup service for its own data protection.

Disaster Recovery

AWS Organizations does not directly support AWS Elastic Disaster Recovery integration as it's an account management service rather than a workload service. The service itself is highly available across AWS regions with automatic failover capabilities. Organizations enables disaster recovery governance by allowing centralized policies and resource sharing across accounts in different regions, supporting multi-region DR architectures for workloads within member accounts.

Recovery Objectives

AWS Organizations helps customers meet RPO and RTO objectives by enabling centralized governance of backup and disaster recovery policies across member accounts. Through AWS Backup integration, customers can enforce consistent backup schedules and retention policies. The service supports multi-region account organization for geographic distribution and enables resource sharing for DR infrastructure. However, specific RPO/RTO targets depend on workload-level configurations within individual accounts rather than the Organizations service itself.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/organizations/latest/userguide/orgs_reference_limits.html

FAQ: https://aws.amazon.com/organizations/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/organizations/latest/userguide/orgs_introduction.html

User Guide: https://docs.aws.amazon.com/organizations/latest/userguide/orgs_getting-started.html

API Reference: https://docs.aws.amazon.com/organizations/latest/APIReference/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://docs.aws.amazon.com/organizations/latest/userguide/pricing.html

Cost Optimization

AWS Organizations provides unique cost optimization through consolidated billing that aggregates usage across all member accounts to reach volume discount tiers faster. Reserved Instances and Savings Plans are shared automatically across the organization, maximizing utilization and savings potential:

- The service enables centralized cost allocation using tags and Cost Categories for departmental chargeback and budget management
- Organizations supports AWS Cost Explorer integration for organization-wide cost analysis and AWS Compute Optimizer for rightsizing recommendations across accounts
- Customers can implement cost control policies through SCPs to prevent expensive resource deployment and achieve economies of scale through centralized procurement and resource sharing via AWS Resource Access Manager

Savings Considerations

AWS Organizations delivers substantial cost savings through several mechanisms. Consolidated billing eliminates individual account fees and provides volume discounts across aggregated usage, with Reserved Instances and Savings Plans automatically shared for maximum utilization:

- The service itself is completely free, adding no cost while enabling significant savings
- Centralized resource sharing through AWS RAM reduces duplicate infrastructure costs, while automated account management reduces operational overhead
- Tag policies enable accurate cost allocation for departmental chargeback and budget optimization
- Organizations enables bulk purchasing of Reserved Instances and Savings Plans with better utilization rates across multiple accounts, potentially saving up to 72% on compute costs compared to individual account management

Service Name

AWS Outposts

Service Overview

AWS Outposts is a fully managed service that extends AWS infrastructure, services, APIs, and tools to customer premises. It enables customers to build and run applications on-premises using the same programming interfaces as in AWS Regions, while using local compute and storage resources for lower latency and local data processing needs. An Outpost is a pool of AWS compute and storage capacity deployed at a customer site, which AWS operates, monitors, and manages. Available in rack and server form factors, Outposts provides consistent hybrid experience with native AWS services including EC2, EBS, S3, RDS, and container services running locally while maintaining connection to the parent AWS Region.

Key Use Cases

- Low-latency applications: Workloads requiring sub-millisecond latencies and real-time processing near end users or on-premises systems
- Data residency and compliance: Applications that must keep data on-premises due to regulatory requirements, data sovereignty, or compliance mandates
- Local data processing: Processing large volumes of data locally before sending insights to the cloud, reducing data transfer costs and improving performance
- Hybrid workloads: Applications with local system interdependencies that need seamless integration between on-premises and cloud resources
- Edge computing: Running compute and storage services at edge locations for IoT, manufacturing, healthcare, and content delivery use cases

Features

- **Compute Services**: Amazon EC2 instances (M5, C5, R5, M7i, C7i, R7i, G4dn), Amazon ECS clusters, Amazon EKS nodes, with support for both general purpose and specialized workloads
- **Storage Services**: Amazon EBS volumes with gp2 storage, Amazon S3 on Outposts buckets, EBS Local Snapshots, and support for third-party block storage integration
- **Database Services**: Amazon RDS DB instances, Amazon DynamoDB, Amazon DocumentDB, and Amazon ElastiCache nodes running locally on Outposts
- **Networking**: VPC extension, local gateway connectivity, Application Load Balancers, customer-owned IP addresses, and direct VPC routing to on-premises networks
- **Container Services**: Amazon ECS, Amazon EKS, AWS Fargate support for containerized applications with local orchestration capabilities

- **Management Tools**: AWS Systems Manager, AWS Config, AWS Resource Access Manager, and standard AWS APIs, CLI, and SDKs for consistent management experience

Value Proposition

AWS Outposts provides unique advantages for organizations requiring on-premises compute while maintaining cloud benefits. Unlike traditional infrastructure, customers receive fully managed AWS services with automatic updates, patches, and maintenance handled by AWS:

- The service offers true hybrid consistency with identical APIs, tools, and services as AWS Regions, eliminating operational complexity
- Key differentiators include ultra-low latency for real-time applications, guaranteed data residency for compliance requirements, seamless integration with existing AWS workloads, and the ability to scale on-premises capacity without managing hardware procurement or lifecycle management
- Organizations avoid upfront infrastructure investments while gaining enterprise-grade security, reliability, and the AWS ecosystem's breadth of services locally

Service Integration

AWS Outposts integrates extensively with core AWS services to provide seamless hybrid functionality. VPC services extend naturally to Outposts with subnets, security groups, and routing tables:

- IAM provides unified identity management across on-premises and cloud resources
- CloudWatch enables monitoring and logging for Outposts resources alongside regional services
- AWS Direct Connect and VPN services establish secure connectivity to parent AWS Regions
- S3 integration supports data transfer between local S3 on Outposts and regional S3 buckets
- Route 53 provides DNS services with local resolvers available on racks
- AWS Backup, DataSync, and Elastic Disaster Recovery enable data protection strategies
- Container services integrate with Amazon ECR for image management, while database services maintain compatibility with AWS database tools and features

Onboarding and Offboarding

Getting started with AWS Outposts involves several key steps managed through the AWS Console. First, customers create an Outpost site defining the physical location and network requirements:

- Next, they create an Outpost configuration specifying compute and storage capacity needs, then place an order selecting payment options and deployment details
- AWS handles delivery, installation, and activation, requiring customer-provided power, cooling, and network connectivity
- During installation, AWS technicians configure the service link to the parent AWS Region and integrate with the local network
- Customers must have appropriate IAM permissions and enterprise support subscription
- After activation, standard AWS APIs, CLI, and console access enables resource provisioning and management identical to cloud operations

Getting Started Guide: https://docs.aws.amazon.com/outposts/latest/userguide/get-started-outposts.html

Data Removal

At term end, customers have options to renew, return, or convert to month-to-month billing. For data removal, customers must terminate all EC2 instances, delete EBS volumes and S3 buckets, remove VPC associations, and destroy the AWS Nitro Security Key (NSK) which cryptographically shreds all data on the Outpost. AWS provides at least 5 business days notice for decommissioning. The retrieval team powers down equipment, and customers can destroy the NSK themselves or have AWS do it during retrieval, ensuring complete data destruction.

Backup/Restore and Disaster Recovery

AWS Outposts supports multiple backup and disaster recovery strategies. EBS snapshots can be stored locally on S3 on Outposts or in the parent AWS Region for redundancy. S3 Versioning and Cross-Region Replication protect object data. AWS Elastic Disaster Recovery continuously replicates instances and storage between Outposts, on-premises systems, and AWS Regions. DataSync facilitates data transfer for backup purposes. Third-party backup solutions from AWS Partners provide additional options for comprehensive data protection strategies.

Backup Capabilities

AWS Outposts provides native backup capabilities through EBS snapshots stored locally or in AWS Regions, S3 object versioning and replication, and integration with AWS Backup service. Local snapshots on S3 on Outposts meet data residency requirements while regional snapshots provide geographic redundancy. The service supports automated backup policies and point-in-time recovery for applications and databases running on Outposts infrastructure.

Disaster Recovery

AWS Outposts supports disaster recovery through AWS Elastic Disaster Recovery (DRS) which replicates instances between Outposts, on-premises systems, and AWS Regions. Multi-AZ deployments across multiple Outposts provide local redundancy:

- Route 53 health checks enable automatic failover
- Application Load Balancers distribute traffic and support failover scenarios
- The service maintains connectivity to parent AWS Region for control plane operations during local failures

Recovery Objectives

AWS Outposts enables customers to meet aggressive RPO and RTO objectives through local data processing and storage capabilities. Local snapshots and replication can achieve RPO measured in minutes, while automatic failover mechanisms and redundant infrastructure support RTO targets of minutes to hours. Multi-rack deployments and regional backup strategies provide multiple recovery tiers to match business requirements and criticality levels for different workloads.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/outposts/latest/userguide/outposts-limits.html

FAQ: https://aws.amazon.com/outposts/rack/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/outposts/latest/userguide/what-is-outposts.html

User Guide: https://docs.aws.amazon.com/outposts/latest/userguide/index.html

API Reference:
https://docs.aws.amazon.com/outposts/latest/APIReference/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/outposts/rack/pricing/

Cost Optimization

AWS Outposts offers several cost optimization approaches unique to hybrid infrastructure. Customers can optimize by right-sizing Outpost configurations to match workload requirements, avoiding over-provisioning:

- Local data processing reduces data transfer costs to AWS Regions
- Choosing appropriate payment options (All Upfront, Partial Upfront, No Upfront) based on cash flow preferences provides cost flexibility
- Shared resource models through AWS Resource Access Manager enable cost sharing across accounts
- Local S3 storage reduces regional storage and transfer costs
- Consolidating multiple workloads on single Outposts maximizes hardware utilization and reduces per-workload costs

Savings Considerations

Primary cost savings come from eliminating traditional data center infrastructure management overhead, including hardware procurement, maintenance contracts, and specialized staff. Reduced data transfer costs through local processing and storage provide ongoing operational savings:

- Avoiding over-provisioning through AWS's managed capacity and right-sizing recommendations optimizes costs
- Three-year term commitments provide significant discounts over month-to-month pricing
- Shared infrastructure costs across multiple workloads and applications improve overall ROI
- Integration with existing AWS tooling and automation reduces operational expenses compared to managing separate on-premises solutions

Service Name

AWS Parallel Computing Service

Service Overview

AWS Parallel Computing Service (AWS PCS) is a managed service that simplifies running and scaling high performance computing (HPC) workloads and building scientific and engineering models on AWS using Slurm. The service integrates AWS compute, storage, networking, and visualization resources while streamlining cluster operations through built-in management and observability capabilities. AWS PCS enables users to focus on research and innovation by providing a familiar environment for running applications and jobs without the operational overhead of managing HPC infrastructure. The service utilizes Slurm for workload management and automatically handles cluster scaling, maintenance, and updates.

Key Use Cases

- Computational chemistry and chemical formulation discovery for pharmaceutical research and materials science
- Scientific simulations including computational fluid dynamics (CFD), structural analysis, and engineering modeling
- Genomic sequencing, medical imaging analysis, and clinical research simulations for healthcare organizations
- Cryogenic electron microscopy (Cryo-EM) workloads for drug discovery and molecular structure determination
- Aerospace simulation workloads including mission planning, space operations, and vehicle innovation
- Multi-node MPI jobs requiring elastic scaling across distributed compute resources

Features

- **Managed Slurm Scheduler**: Fully managed Slurm workload manager with automated scaling and maintenance
- **Elastic Compute Node Groups**: Automatically scaling EC2 instances with customizable instance types and purchase options
- **Queue Management**: Job scheduling and execution across compute node groups with customizable policies
- **Built-in Observability**: Job completion logs, monitoring, and cluster management through AWS console
- **Slurm Customization**: Over 60 configurable Slurm parameters for scheduling, resource allocation, and access control

- **Managed Accounting**: Slurm accounting feature for usage tracking, resource limits, and chargeback reporting
- **Capacity Blocks Integration**: Support for EC2 Capacity Blocks to reserve GPU instances up to 8 weeks in advance
- **REST API Support**: HTTP interface for programmatic interaction with Slurm workload manager

Value Proposition

AWS PCS offers significant advantages over traditional on-premises HPC solutions and self-managed alternatives like AWS ParallelCluster. The service eliminates operational overhead by providing fully managed infrastructure with automatic scaling, patching, and updates:

- Organizations benefit from faster time-to-science with instant access to virtually unlimited computational resources, avoiding long queue times typical of traditional HPC environments
- AWS PCS provides enterprise-grade reliability with full AWS support and service-level agreements, making it ideal for production workloads requiring uptime guarantees
- The service offers cost optimization through elastic scaling, spot instance support, and pay-as-you-go pricing, eliminating the need for large capital investments in depreciating hardware
- Integration with the broader AWS ecosystem enables seamless access to storage, networking, and visualization tools

Service Integration

AWS PCS integrates deeply with core AWS services to provide comprehensive HPC solutions. Amazon EC2 provides the underlying compute infrastructure with support for various instance types including GPU-accelerated instances:

- Amazon EFS and Amazon FSx for Lustre deliver shared storage solutions for home directories and high-performance workloads
- Amazon VPC enables secure networking with customizable subnets and security groups
- AWS IAM provides identity and access management for cluster resources and user permissions
- Amazon CloudWatch delivers monitoring and logging capabilities including job completion logs
- AWS Systems Manager enables administrative access and management tasks
- AWS Secrets Manager handles cluster secret rotation for secure communications
- The service also integrates with AWS Backup for data protection and AWS PrivateLink for secure private connectivity

Onboarding and Offboarding

Customers begin by completing prerequisites including AWS account setup, AWS CLI installation, and required IAM permissions. The getting started process involves creating a VPC with appropriate subnets, configuring security groups, and setting up shared storage using Amazon EFS and FSx for Lustre:

- Users then create an AWS PCS cluster specifying the Slurm scheduler configuration, followed by defining compute node groups with desired instance types and scaling parameters
- A queue must be created to manage job scheduling across compute resources
- AWS provides comprehensive tutorials and CloudFormation templates to automate cluster deployment
- The service offers both console-based and CLI-driven setup options
- Once configured, users can connect via SSH or AWS Systems Manager to submit and manage jobs using standard Slurm commands

Getting Started Guide: https://docs.aws.amazon.com/pcs/latest/userguide/getting-started.html

Data Removal

When offboarding from AWS PCS, customers can delete cluster resources through the AWS console or CLI, which automatically terminates all associated EC2 instances and removes cluster configurations. Data stored in customer-owned resources like EFS and FSx file systems must be manually deleted by the customer. Job completion logs and accounting data stored in AWS-managed databases are automatically removed when clusters are deleted. Customers should backup any important data before initiating the deletion process as it is irreversible.

Backup/Restore and Disaster Recovery

AWS PCS leverages integrated AWS services for backup and disaster recovery capabilities. Customer data stored in Amazon EFS and FSx for Lustre can be backed up using AWS Backup for automated, policy-driven protection. EC2 instances can be protected through EBS snapshots and AMI creation. Job completion logs are automatically stored in CloudWatch Logs or S3 for durability. The service supports cross-region deployment for disaster recovery scenarios, though cluster configurations must be recreated in alternate regions.

Backup Capabilities

AWS PCS does not provide native backup capabilities as a service but integrates with AWS Backup for comprehensive data protection. Customer workload data stored on shared file systems (EFS, FSx) can be automatically backed up using AWS Backup policies:

- Compute node data can be protected through EBS snapshots and custom AMI creation
- Job metadata and logs are durably stored in AWS-managed services with built-in redundancy

Disaster Recovery

AWS PCS supports disaster recovery through multi-region deployment strategies and integration with AWS disaster recovery services. While clusters cannot be directly replicated across regions, customers can deploy identical cluster configurations in multiple regions using CloudFormation templates:

- Data replication can be achieved through EFS and FSx cross-region backup and restore capabilities
- AWS Elastic Disaster Recovery can protect compute workloads and data

Recovery Objectives

AWS PCS helps customers meet RPO and RTO objectives through automated scaling and rapid cluster deployment capabilities. New clusters can be provisioned within minutes using pre-configured templates, supporting aggressive RTO requirements. Data RPO objectives are met through integration with AWS Backup for frequent, automated snapshots of shared storage. The service's elastic nature allows rapid scaling to handle recovery workloads and catch-up processing after incidents.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/pcs/latest/userguide/service-endpoints-quotas.html

FAQ: https://aws.amazon.com/pcs/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/pcs/latest/userguide/what-is-service.html

User Guide: https://docs.aws.amazon.com/pcs/latest/userguide/getting-started.html

API Reference: https://docs.aws.amazon.com/pcs/latest/APIReference/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/pcs/pricing/

Cost Optimization

AWS PCS offers several unique cost optimization features including elastic auto-scaling that automatically adds and removes compute nodes based on job queue demand, eliminating idle resource costs. The service supports EC2 Spot instances for significant cost savings on fault-tolerant workloads, and mixed instance types within compute node groups for optimal price-performance ratios:

- Capacity Blocks integration enables reserving GPU instances up to 8 weeks in advance with discounted rates
- The managed accounting feature provides detailed usage tracking for chargeback and resource optimization
- Pay-as-you-go pricing eliminates capital expenditure requirements, and the service automatically terminates unused resources to prevent unnecessary charges

Savings Considerations

Primary cost savings come from eliminating on-premises HPC infrastructure capital costs and operational overhead. Organizations save on hardware depreciation cycles typically lasting 5-8 years and reduce staffing costs for cluster management and maintenance:

- The elastic scaling model ensures customers only pay for compute resources during active job execution, avoiding the high utilization costs of traditional HPC systems
- Spot instance support can reduce compute costs by up to 90% for appropriate workloads
- The managed service eliminates costs associated with Slurm expertise, system administration, and infrastructure troubleshooting
- Multi-tenancy capabilities allow sharing resources across projects and teams, improving overall utilization efficiency and reducing per-user costs

Service Name

AWS Payment Cryptography

Service Overview

AWS Payment Cryptography is a managed service that provides access to cryptographic functions and key management for payment processing in compliance with PCI standards without requiring dedicated payment HSM instances. The service enables payment functions like acquirers, payment facilitators, networks, switches, processors, and banks to move cryptographic operations closer to cloud applications while minimizing dependencies on auxiliary data centers. It leverages PCI PTS HSM V3 and FIPS 140-2 Level 3 certified hardware with low latency and high availability, eliminating operational requirements of on-premises HSMs such as hardware provisioning, key material management, and emergency backup maintenance.

Key Use Cases

- Card issuing operations: EMV PIN change, cardholder-selectable PINs, PIN reveal, CVV/CVV2 generation and validation, and ARQC processing for transaction authorization
- Acquiring and payment facilitation: Data decryption from payment terminals using TDES, AES, and DUKPT techniques with PCI P2PE compliance, PIN translation between encryption keys, and MAC verification for data integrity
- Key exchange and synchronization: Electronic key sharing with partners using TR-34, TR-31, and ECDH protocols, eliminating manual paper-based key component sharing and enabling secure migration from on-premises HSMs

Features

- **Key Management**: Create, manage, import, and export symmetric and asymmetric keys (TDES, AES, RSA) with alias management, tagging, and TR-31/TR-34 support
- **Cryptographic Operations**: Encrypt, decrypt, re-encrypt data, generate/validate CVV/CVV2/ARQC, PIN translation, MAC generation/verification, and cardholder data processing
- **Multi-Region Replication**: Automatic key replication across AWS regions for high availability, disaster recovery, and global operations with consistent key IDs
- **PCI Compliance**: Built-in compliance with PCI PIN, PCI P2PE, and PCI DSS standards using PCI PTS HSM V3 certified hardware
- **Electronic Key Exchange**: Support for TR-34, TR-31, and ECDH protocols for secure electronic key exchange with partners and HSM synchronization
- **Post-Quantum Security**: Hybrid post-quantum TLS support using ML-KEM cryptography for securing API calls against future quantum computing threats

Value Proposition

AWS Payment Cryptography eliminates the operational complexity and costs of managing on-premises payment HSMs while providing elastic scalability and built-in PCI compliance. The service offers significant advantages over traditional solutions including elimination of hardware provisioning, maintenance, and emergency backup requirements in secure facilities:

- Customers benefit from industry-leading availability SLAs, automatic failover between availability zones, and the ability to electronically share keys with partners instead of using paper-based processes
- The service provides cost optimization through tiered pricing models and reduces total cost of ownership compared to dedicated HSM infrastructure
- Integration with AWS services enables modern cloud-native payment architectures with API-based operations and faster time-to-market for payment applications

Service Integration

AWS Payment Cryptography integrates deeply with core AWS security and monitoring services for comprehensive payment processing solutions. Primary integrations include AWS Identity and Access Management (IAM) for authentication, authorization, and access control policies with support for dual control operations:

- AWS CloudTrail provides comprehensive audit logging for all key management and cryptographic operations, while Amazon CloudWatch enables monitoring and alerting
- The service supports Amazon VPC and AWS PrivateLink for secure network connectivity, and AWS Direct Connect for dedicated network connections
- Integration with AWS Key Management Service (KMS) and AWS CloudHSM provides additional cryptographic options for complementary use cases, enabling customers to build complete payment security architectures across multiple AWS security services

Onboarding and Offboarding

Customers get started by ensuring proper IAM permissions and installing the AWS CLI, SDKs, or using AWS APIs for service access. The onboarding process begins with creating cryptographic keys using the Control Plane API, specifying key attributes like algorithm (TDES, AES, RSA), usage (CVV generation, PIN translation, etc.), and key modes of use:

- The Getting Started tutorial demonstrates creating a CVK key for CVV2 operations, followed by cryptographic operations using the Data Plane API
- AWS provides comprehensive workshops, GitHub samples, and documentation to help customers explore deployment patterns and common use cases

- Customers can import existing keys from on-premises HSMs using TR-31 and TR-34 protocols or create new keys within the service, with immediate access to cryptographic operations upon successful key creation

Getting Started Guide: https://docs.aws.amazon.com/payment-cryptography/latest/userguide/getting-started.html

Data Removal

Key deletion in AWS Payment Cryptography involves scheduling keys for deletion with a configurable waiting period (3-180 days, default 7 days) during which keys enter DELETE_PENDING state and cannot be used for cryptographic operations. Deletion is irreversible and permanently removes key material and metadata, making decryption of previously encrypted data impossible. Customers can cancel deletion before the scheduled time using the restore-key operation. For Multi-Region keys, deleting a Primary Region key automatically deletes associated Replica Region keys across all replication regions, ensuring complete data removal.

Backup/Restore and Disaster Recovery

AWS Payment Cryptography provides built-in high availability and disaster recovery through Multi-Region key replication, automatically distributing keys across AWS regions with eventual consistency. The service maintains durability matching AWS's highest durability services and operates across at least three availability zones per region with automatic failover. Key material is protected by HSM AES 256 main keys and stored in encrypted databases with replication to in-memory databases on Payment Cryptography servers, ensuring resilience against hardware failures and regional outages.

Backup Capabilities

AWS Payment Cryptography does not integrate with AWS Backup as it manages key material internally through HSM-based protection and encrypted storage. The service provides inherent backup through Multi-Region replication capabilities, cross-region key synchronization using TR-34 and TR-31 protocols, and durable key storage protected by regional and profile main keys. Key durability is designed to match AWS's highest durability services with automatic replication across availability zones within each region.

Disaster Recovery

The service does not integrate with AWS Elastic Disaster Recovery but provides native disaster recovery through Multi-Region key replication and regional isolation design. AWS Payment Cryptography supports automatic failover between availability zones with zero planned downtime and cross-region key distribution for business continuity. The service enables customers to implement custom disaster recovery solutions using the cross-region replication capabilities and industry-standard key exchange protocols for maintaining operational resilience.

Recovery Objectives

AWS Payment Cryptography helps customers achieve aggressive RPO and RTO objectives through Multi-Region key replication with near real-time synchronization and automatic availability zone failover. The service's regional isolation design and elastic scaling capabilities enable customers to maintain cryptographic operations during regional outages by utilizing replicated keys in secondary regions. High availability SLAs and low-latency operations support critical payment processing requirements with minimal recovery time objectives for business-critical payment functions.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/payment-cryptography/latest/userguide/quotas.html

FAQ: https://aws.amazon.com/payment-cryptography/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/payment-cryptography/latest/userguide/what-is.html

User Guide: https://docs.aws.amazon.com/payment-cryptography/latest/userguide/

API Reference: https://docs.aws.amazon.com/payment-cryptography/latest/APIReference/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/payment-cryptography/pricing/

Cost Optimization

AWS Payment Cryptography offers unique cost optimization through recent pricing reductions of up to 63% for API requests and introduction of tiered key pricing models that scale with volume. The service eliminates traditional HSM hardware costs including procurement, maintenance, facility requirements, and dedicated personnel:

- Customers benefit from elastic scaling that charges only for actual usage rather than provisioned capacity, and Multi-Region replication reduces the need for separate disaster recovery infrastructure
- The tiered pricing structure enables cost-effective scaling for high-volume cryptographic operations while the managed service model eliminates operational overhead costs associated with on-premises HSM management and PCI compliance maintenance

Savings Considerations

Primary cost savings derive from eliminating dedicated payment HSM infrastructure, including hardware acquisition, data center colocation, secure facility requirements, and specialized technical personnel. The managed service model removes ongoing operational costs such as HSM maintenance, firmware updates, emergency backup management, and PCI compliance auditing overhead:

- Customers save on disaster recovery infrastructure through built-in Multi-Region replication and high availability features
- Electronic key exchange capabilities eliminate costs associated with physical key component transportation and manual key management processes
- The elastic pricing model provides significant savings compared to fixed HSM capacity provisioning, enabling customers to optimize costs based on actual transaction volumes and seasonal payment processing patterns

Service Name

AWS PrivateLink

Service Overview

AWS PrivateLink is a highly available, scalable technology that enables private connectivity between VPCs, AWS services, and on-premises networks without using the public internet. It creates secure connections using private IP addresses within Amazon's network infrastructure, eliminating the need for internet gateways, NAT devices, public IP addresses, Direct Connect connections, or VPN connections. Users control specific API endpoints, sites, services, and resources that are reachable from their VPC, significantly reducing attack surface and enhancing security. Traffic remains entirely within the AWS network backbone, providing improved performance, security, and simplified network architecture for accessing AWS services, third-party SaaS applications, and custom services across VPCs and accounts.

Key Use Cases

- SaaS Application Access: Privately access third-party SaaS applications from VPCs without public internet exposure or complex network configurations
- Shared Services Architecture: Enable common services in shared VPCs to be consumed by workloads in separate VPCs without VPC peering limitations
- Hybrid Cloud Integration: Extend services to on-premises resources by connecting endpoint services to data center resources via Network Load Balancers
- Microservices Communication: Facilitate secure inter-service communication where teams create endpoint services for their microservices and allow controlled access by other teams
- Multi-Region Service Access: Provide cross-region access to services while maintaining private communication over Amazon's network infrastructure

Features

- **Interface VPC Endpoints**: Elastic network interfaces that provide private connectivity to AWS services, third-party services, and VPC endpoint services
- **VPC Endpoint Services**: Allow service providers to expose their applications behind Network Load Balancers for private consumption by other AWS customers
- **Resource VPC Endpoints**: Direct private connectivity to specific VPC resources like databases and clusters without requiring load balancing
- **Cross-Region Connectivity**: Extends private connectivity to AWS services across regions without internet or inter-region peering
- **Security Group Integration**: Apply VPC security groups to control access at the network interface level for granular security policies

- **Endpoint Policies**: Fine-grained access control policies to restrict access to specific AWS principals, actions, and resources
- **DNS Resolution**: Private DNS names for endpoints enabling seamless application integration without code changes
- **High Availability**: Horizontally scalable and redundant across multiple Availability Zones for fault tolerance

Value Proposition

AWS PrivateLink offers superior security by keeping all traffic within Amazon's private network, eliminating internet exposure and reducing attack surface. It provides significant cost savings by avoiding NAT gateway charges and internet data transfer costs, with potential reductions of up to 92% compared to traditional internet-based connectivity:

- The service simplifies network architecture by removing the need for complex VPC peering, internet gateways, or VPN connections while supporting overlapping IP address ranges between VPCs
- PrivateLink delivers consistent low-latency performance through Amazon's high-speed backbone network and scales automatically to handle millions of requests per second
- It enables compliance with strict data residency and security requirements while facilitating seamless cloud migration and hybrid architectures

Service Integration

AWS PrivateLink integrates with over 100 AWS services including Amazon S3, DynamoDB, Lambda, EC2, RDS, CloudWatch, IAM, and many others through interface endpoints. It works seamlessly with AWS Direct Connect and Site-to-Site VPN for hybrid connectivity, enabling on-premises access to PrivateLink endpoints:

- The service integrates with Amazon VPC security groups, NACLs, and route tables for comprehensive network security
- It supports AWS CloudFormation for infrastructure-as-code deployments and integrates with AWS CloudTrail for audit logging
- PrivateLink works with Elastic Load Balancing (specifically Network Load Balancers) to publish services and connects with Amazon VPC Lattice for service mesh architectures
- It also integrates with AWS Resource Access Manager for cross-account resource sharing

Onboarding and Offboarding

Customers begin by creating VPC endpoints through the AWS Management Console, AWS CLI, or CloudFormation templates to connect to desired services. For consuming AWS

services, select the service from available options and specify target subnets and security groups:

- For accessing custom services, coordinate with service providers to obtain endpoint service names and ensure proper allowlisting
- Service providers create endpoint services by setting up Network Load Balancers and configuring VPC endpoint services with appropriate acceptance settings
- The process includes configuring DNS settings, security groups, and endpoint policies for access control
- Testing connectivity and monitoring through CloudWatch metrics ensures proper functionality
- Documentation provides step-by-step tutorials for common scenarios including connecting to CloudWatch, S3, and custom applications

Getting Started Guide: https://docs.aws.amazon.com/vpc/latest/privatelink/getting-started.html

Data Removal

Offboarding involves deleting VPC endpoints through the AWS Management Console, CLI, or API, which immediately terminates private connectivity. Service providers must delete endpoint services and associated Network Load Balancer configurations. All endpoint policies, security group rules, and DNS configurations should be cleaned up. VPC endpoint connections are automatically removed when endpoints are deleted, and no residual data remains within the PrivateLink service itself as it only facilitates connectivity.

Backup/Restore and Disaster Recovery

AWS PrivateLink itself does not provide backup or disaster recovery capabilities as it is a networking service that facilitates connectivity. However, it enables private access to AWS Backup and AWS Elastic Disaster Recovery services, allowing customers to implement backup and DR strategies while maintaining private network connectivity. The service supports multi-AZ deployments for high availability and can be configured across regions for disaster recovery architectures.

Backup Capabilities

PrivateLink does not have inherent backup capabilities as it is a networking connectivity service. Configuration settings for VPC endpoints and endpoint services can be captured through AWS Config and replicated via CloudFormation templates. The service enables private connectivity to AWS Backup service, allowing customers to implement backup strategies while maintaining network isolation and security.

Disaster Recovery

PrivateLink supports disaster recovery by enabling cross-region connectivity and multi-AZ deployments for high availability. While not directly integrated with AWS Elastic Disaster Recovery, it provides the private network connectivity required for DR scenarios. Customers can configure PrivateLink endpoints in multiple regions and leverage the service's inherent redundancy across Availability Zones for resilient architectures.

Recovery Objectives

PrivateLink supports RPO and RTO objectives by providing highly available, low-latency connectivity with automatic scaling and redundancy across multiple Availability Zones. The service enables rapid failover scenarios through DNS-based routing and supports cross-region connectivity for disaster recovery. By eliminating internet dependencies and providing consistent private network performance, it helps maintain predictable recovery timeframes and minimizes data exposure during recovery operations.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/vpc/latest/privatelink/vpc-limits-endpoints.html

FAQ: https://aws.amazon.com/privatelink/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/vpc/latest/privatelink/what-is-privatelink.html

User Guide: https://docs.aws.amazon.com/vpc/latest/privatelink/getting-started.html

API Reference: https://docs.aws.amazon.com/AWSEC2/latest/APIReference/operation-list-privatelink.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/privatelink/pricing/

Cost Optimization

PrivateLink offers significant cost optimization by eliminating NAT gateway charges and reducing internet data transfer costs, with potential savings up to 92% compared to traditional internet-based connectivity. Customers can optimize costs by consolidating multiple service connections through single VPC endpoints where possible:

- Resource endpoints provide cost-effective access to databases and specific resources without requiring load balancers
- Cross-region connectivity should be evaluated against data transfer costs versus security and compliance benefits
- Proper endpoint policy configuration prevents unnecessary data processing charges by restricting access to only required resources and principals

Savings Considerations

Primary cost savings include elimination of NAT gateway hourly charges and internet data transfer fees, particularly significant for high-volume data processing workloads. PrivateLink's per-GB processing fees are often lower than equivalent internet data transfer and NAT gateway costs:

- Organizations can reduce network infrastructure complexity and associated management costs
- Cross-region private connectivity eliminates the need for expensive dedicated connections while maintaining security
- For SaaS integrations, private connectivity reduces bandwidth costs and improves predictable billing
- Shared endpoint services can distribute costs across multiple consumers, making enterprise services more cost-effective

Service Name

AWS Resilience Hub

Service Overview

AWS Resilience Hub is a centralized service that enables organizations to manage, assess, and improve the resilience posture of their applications running on AWS. The service provides a comprehensive platform for defining resilience goals, continuously assessing application weaknesses against Recovery Time Objectives (RTO) and Recovery Point Objectives (RPO), and implementing recommendations based on AWS Well-Architected Framework best practices. AWS Resilience Hub integrates with AWS Fault Injection Service to conduct chaos engineering experiments, automatically discovers and maps application resources, and provides actionable insights through drift detection, resiliency scoring, and operational recommendations. The service helps organizations transform resilience management from a reactive process into a proactive strategy by providing automated assessments, detailed recommendations, and clear resilience metrics across multi-account and multi-region architectures.

Key Use Cases

- Application Resilience Assessment: Continuously assess applications against defined RTO/RPO targets to identify weaknesses and ensure compliance with resilience policies across different failure scenarios including application, infrastructure, AZ, and regional outages
- Proactive Disaster Recovery Planning: Define and validate disaster recovery strategies by simulating various disruption scenarios using AWS FIS experiments and implementing standard operating procedures (SOPs) for efficient recovery
- Multi-Account Resilience Management: Centrally manage resilience across distributed, multi-account cloud architectures using a hub-and-spoke model to ensure consistent resilience policies and assessments
- CI/CD Pipeline Integration: Integrate resilience validation into software development lifecycle to identify and resolve resilience weaknesses before applications are released into production
- Operational Readiness and Compliance: Maintain regulatory and compliance requirements by tracking resilience posture, implementing recommended alarms, and ensuring applications meet defined business continuity objectives

Features

- **Resilience Dashboard**: Visual dashboard providing comprehensive overview of application resilience status, compliance, drift detection, and resiliency scores across multiple applications

- **Application Discovery and Management**: Automatic discovery and mapping of AWS resources from CloudFormation stacks, Resource Groups, EKS clusters, and Terraform state files into logical application components
- **Resiliency Policy Management**: Define and manage customizable resilience policies with specific RTO and RPO targets for different disruption types including software, hardware, AZ, and regional failures
- **Automated Assessments**: Continuous resilience assessments with scheduled daily evaluations, drift detection capabilities, and compliance reporting against defined policies
- **AWS FIS Integration**: Integrated chaos engineering capabilities with tailored fault injection experiments to validate application resilience and verify recovery procedures
- **Operational Recommendations**: AI-powered recommendations for infrastructure improvements, alarm configurations, and standard operating procedures based on AWS Well-Architected Framework
- **Multi-Account Support**: Hub-and-spoke architecture enabling centralized resilience management across multiple AWS accounts with cross-account role assumptions and trust relationships
- **Resiliency Scoring**: Quantitative scoring system measuring application readiness to handle disruptions based on adherence to recommendations, alarms, SOPs, and successful tests

Value Proposition

AWS Resilience Hub provides unique value through its comprehensive, centralized approach to application resilience management that transforms reactive disaster recovery into proactive resilience engineering. Unlike fragmented solutions, it offers unified visibility across multi-account architectures with automated resource discovery and continuous assessment capabilities:

- The service uniquely integrates with AWS Well-Architected Framework to provide evidence-based recommendations and incorporates AWS FIS for validated chaos engineering testing
- Customers benefit from AI-powered insights using Amazon Bedrock for assessment summaries, cost-optimized resilience recommendations with multiple optimization strategies (cost, minimal changes, best AZ/region recovery), and seamless integration with existing CI/CD pipelines
- The platform reduces operational overhead through automated drift detection, scheduled assessments, and standardized operating procedures while ensuring regulatory compliance through detailed reporting and metrics
- Organizations can achieve faster time-to-resilience with pre-built policies, experiment templates, and actionable recommendations that eliminate guesswork in disaster recovery planning

Service Integration

AWS Resilience Hub deeply integrates with core AWS services to provide comprehensive resilience management. It works natively with AWS CloudFormation and Terraform for infrastructure-as-code resource discovery, Amazon EKS and ECS for containerized workload assessment, and AWS Resource Groups for logical resource organization:

- The service integrates with AWS Fault Injection Service for chaos engineering experiments and AWS Systems Manager for standard operating procedures automation
- It leverages Amazon CloudWatch for alarm management and monitoring, AWS SNS for drift notifications, and Amazon S3 for metrics export and storage
- The platform connects with AWS Backup and AWS Elastic Disaster Recovery services for backup and disaster recovery capabilities assessment
- Integration with AWS Trusted Advisor provides additional resilience checks, while AWS CloudTrail enables audit logging of resilience activities
- For enterprise deployments, it works with AWS Organizations through IAM cross-account roles and supports myApplications for unified application management
- The service also integrates with Amazon Bedrock for AI-powered assessment summaries and recommendations, providing intelligent insights across the entire AWS ecosystem

Onboarding and Offboarding

Getting started with AWS Resilience Hub requires setting up IAM permissions and roles, particularly the AWSResilienceHubAssessmentRole for resource access. Customers begin by adding applications through resource collections from CloudFormation stacks, AWS Resource Groups, EKS clusters, or Terraform state files:

- The process involves creating or selecting resiliency policies with appropriate RTO/RPO targets, publishing the application version, and running initial assessments
- For multi-account deployments, organizations establish cross-account trust relationships using AWSResilienceHubCrossAccountRole
- The service offers a six-month free trial for the first three applications, allowing customers to evaluate capabilities before committing
- Setup includes configuring SNS topics for drift notifications, scheduled assessments, and implementing recommended improvements
- AWS provides free online training, self-guided labs, GitHub repositories with implementation examples, and comprehensive documentation to accelerate adoption
- Customers can leverage pre-built resiliency policies and experiment templates to quickly establish baseline resilience management

Getting Started Guide: https://docs.aws.amazon.com/resilience-hub/latest/userguide/getting-started.html

Data Removal

When offboarding from AWS Resilience Hub, customers can delete applications, assessments, and associated data through the console or API operations. Deleting an application removes all versions, assessments, and associated metadata from the service. The DeleteApp API operation permanently removes applications and cannot be undone. Customers should first export any required metrics or reports before deletion. Assessment data, recommendation templates, and past assessments are retained for 365 days by default before automatic deletion. To completely remove data, customers must explicitly delete individual applications, policies, and exported S3 buckets created by the service. The service automatically cleans up resources and stops charging when applications are deleted.

Backup/Restore and Disaster Recovery

AWS Resilience Hub assesses and recommends backup and disaster recovery capabilities by evaluating integration with AWS Backup and AWS Elastic Disaster Recovery services. The service checks whether applications have appropriate backup plans, cross-region replication, and disaster recovery configurations. It provides recommendations for implementing backup strategies across supported resources like RDS, EFS, EBS, and DynamoDB. The platform evaluates disaster recovery readiness across different failure scenarios including AZ and regional outages, recommending multi-AZ deployments, cross-region replication, and standby resources to meet defined RTO/RPO targets.

Backup Capabilities

AWS Resilience Hub itself does not provide backup capabilities but extensively evaluates and recommends AWS Backup integration for applications. The service assesses whether resources have appropriate AWS Backup plans configured and provides recommendations for backup coverage gaps:

- It analyzes backup configurations for services like Amazon RDS, EFS, EBS, DynamoDB, and other supported resources to ensure they align with defined RPO requirements
- The platform generates specific recommendations for implementing backup strategies that meet resilience policy objectives

Disaster Recovery

AWS Resilience Hub supports disaster recovery planning by assessing application readiness for various failure scenarios and integrating with AWS Elastic Disaster Recovery service. The platform evaluates cross-region disaster recovery configurations, multi-AZ deployments, and standby resource availability:

- It provides specific recommendations for implementing disaster recovery strategies including warm standby, pilot light, and backup-and-restore approaches
- The service assesses recovery capabilities against defined RTO targets and recommends improvements to meet disaster recovery objectives across different disruption types

Recovery Objectives

AWS Resilience Hub helps customers meet RPO and RTO objectives by providing continuous assessment against defined targets, identifying gaps in current configurations, and recommending specific improvements. The service evaluates actual recovery capabilities across different failure scenarios (application, infrastructure, AZ, regional) and provides cost-optimized recommendations to achieve target objectives. It offers multiple optimization strategies including best AZ recovery, best region recovery, least cost, and minimal changes approaches. The platform validates recovery procedures through chaos engineering experiments and provides quantified estimates of achievable RTO/RPO based on current configurations.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/general/latest/gr/resiliencehub.html

FAQ: https://aws.amazon.com/resilience-hub/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/resilience-hub/latest/userguide/what-is.html

User Guide: https://docs.aws.amazon.com/resilience-hub/latest/userguide/

API Reference: https://docs.aws.amazon.com/resilience-hub/latest/APIReference/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/resilience-hub/pricing/

Cost Optimization

AWS Resilience Hub offers unique cost optimization through its multi-dimensional recommendation engine that provides four distinct optimization strategies: cost optimization, minimal changes, best AZ recovery, and best region recovery. The service analyzes current configurations and recommends the most cost-effective approaches to

achieve resilience targets, including right-sizing resources, optimizing backup strategies, and selecting appropriate disaster recovery patterns:

- Customers benefit from the six-month free trial for first three applications, followed by a simple $15 per application per month pricing model with no upfront costs
- The platform helps avoid over-provisioning by providing precise recommendations based on actual RTO/RPO requirements rather than generic disaster recovery approaches
- Cost transparency is built-in with detailed cost estimates for recommended changes, enabling informed decision-making about resilience investments
- The service prevents costly outages by identifying weaknesses before they impact production, providing significant ROI through proactive resilience management

Savings Considerations

Key cost savings with AWS Resilience Hub include avoiding expensive downtime through proactive weakness identification and prevention of over-engineered disaster recovery solutions through precise, requirements-based recommendations. The service eliminates the need for separate resilience management tools and reduces operational overhead through automated assessments and drift detection:

- Organizations save on compliance and audit costs through built-in reporting and documentation capabilities
- The platform's integration with existing AWS services eliminates additional tooling costs while providing comprehensive resilience management
- Customers benefit from optimized resource utilization recommendations that reduce waste in backup and disaster recovery configurations
- The predictable pricing model ($15 per application per month after free trial) enables accurate budgeting compared to complex, usage-based alternatives
- Early identification of resilience gaps prevents costly emergency remediation and reduces the risk of business disruption costs
- Multi-account management capabilities reduce administrative overhead and ensure consistent, efficient resilience policies across organizations

Service Name

AWS Resource Access Manager

Service Overview

AWS Resource Access Manager (RAM) is a service that enables secure sharing of AWS resources across accounts, organizations, and IAM roles and users for supported resource types. It reduces operational overhead by allowing the creation of a resource once and sharing it with other accounts, providing security and consistency by managing access through a single set of policies and permissions, and offering visibility and auditability through integration with Amazon CloudWatch and AWS CloudTrail. RAM simplifies cross-account access by allowing sharing with organizations, organizational units (OUs), and individual accounts without the need to enumerate every account ID. Shared resources appear as native resources in the receiving accounts, and the resource owners can see which principals have access to each resource.

Key Use Cases

- Multi-account resource sharing: Enable organizations to share AWS resources like Transit Gateways, subnets, and Route 53 Resolver rules across multiple AWS accounts without duplicating resources
- Centralized infrastructure management: Create resources once in a central account and share them with other accounts, reducing operational overhead and maintaining consistency across the organization
- Cross-account network connectivity: Share VPC resources, Transit Gateways, and networking components to establish secure connectivity between accounts while maintaining centralized control

Features

- **Resource Sharing**: Share AWS resources across accounts, organizations, organizational units, and IAM roles/users
- **Managed Permissions**: Control access with AWS managed permissions or create custom customer managed permissions
- **Resource-based Policies**: Automatically generate and manage resource-based policies for shared resources
- **Cross-account Visibility**: View shared resources as native resources in consuming accounts with full visibility of access grants
- **Organizations Integration**: Seamless integration with AWS Organizations for organization-wide resource sharing
- **Regional and Global Resources**: Support for sharing both regional and global AWS resources

- **Invitation Management**: Handle resource share invitations for accounts outside the organization
- **Audit and Monitoring**: Integration with CloudTrail and CloudWatch for logging and monitoring resource sharing activities

Value Proposition

AWS Resource Access Manager eliminates the need to provision and manage duplicate resources across multiple AWS accounts, significantly reducing operational overhead and costs. It provides centralized control and governance while maintaining security through managed permissions and resource-based policies:

- RAM simplifies multi-account architectures by enabling seamless resource sharing within AWS Organizations without complex cross-account configurations
- The service offers native integration with AWS services, ensuring shared resources appear and function as native resources in consuming accounts
- Key advantages include no additional charges for the service itself, compliance with major security standards (PCI DSS, FedRAMP, SOC, ISO), and comprehensive audit capabilities through CloudTrail integration

Service Integration

AWS Resource Access Manager integrates deeply with AWS Organizations for organization-wide resource sharing and automatic service-linked role creation. It works seamlessly with AWS Identity and Access Management (IAM) for permission management and access control:

- RAM integrates with Amazon CloudWatch and AWS CloudTrail for comprehensive monitoring, logging, and audit capabilities of all resource sharing activities
- The service supports sharing of resources from over 40 AWS services including Amazon VPC, AWS Transit Gateway, Amazon Route 53 Resolver, AWS License Manager, Amazon Aurora, AWS CodeBuild, Amazon S3, Amazon DynamoDB, and many others
- RAM also integrates with AWS Systems Manager, AWS Private Certificate Authority, AWS Network Firewall, and AWS Backup for specialized sharing scenarios

Onboarding and Offboarding

Getting started with AWS Resource Access Manager is straightforward through the AWS RAM console, AWS CLI, or AWS SDKs. Customers can create resource shares by specifying resources to share, selecting managed permissions for each resource type, and defining the principals (accounts, OUs, or IAM roles/users) to grant access:

- For organization-wide sharing, enable sharing within AWS Organizations first

- The service automatically creates necessary service-linked roles and resource-based policies
- Customers can view, update, and manage resource shares through the console or APIs
- For accounts outside the organization, recipients receive email invitations that must be accepted to gain access to shared resources

Getting Started Guide: https://docs.aws.amazon.com/ram/latest/userguide/getting-started.html

Data Removal

To offboard from AWS Resource Access Manager, customers can stop sharing resources by removing them from resource shares or deleting the resource shares entirely. When a resource share is deleted, all principals lose access to the shared resources immediately. Customers can leave resource shares they no longer need access to, which revokes their access to those shared resources. The resource owner retains full ownership and control of the original resources throughout the sharing process.

Backup/Restore and Disaster Recovery

AWS Resource Access Manager itself does not provide backup and disaster recovery capabilities as it is a resource sharing service. However, RAM can be used to share backup-related resources like AWS Backup vaults, Amazon Aurora clusters, and other resources that support backup functionality. The service enables sharing of disaster recovery resources across accounts, facilitating multi-account disaster recovery strategies.

Backup Capabilities

AWS Resource Access Manager does not have inherent backup capabilities as it is a resource sharing service. RAM does not work directly with AWS Backup for its own data, but it can be used to share AWS Backup resources such as backup vaults and recovery points across accounts, enabling centralized backup management strategies.

Disaster Recovery

AWS Resource Access Manager does not provide disaster recovery capabilities itself, nor does it work with AWS Elastic Disaster Recovery directly. However, RAM can facilitate disaster recovery strategies by enabling the sharing of disaster recovery resources across accounts and regions, supporting multi-account disaster recovery architectures.

Recovery Objectives

AWS Resource Access Manager can indirectly support RPO and RTO objectives by enabling efficient resource sharing across accounts for disaster recovery scenarios. By eliminating the need to duplicate resources across accounts, RAM can help maintain

consistent disaster recovery configurations and reduce recovery complexity, potentially improving RTO through simplified resource access patterns.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/ram/latest/userguide/service-quotas.html

FAQ: https://aws.amazon.com/ram/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/ram/latest/userguide/what-is.html

User Guide: https://docs.aws.amazon.com/ram/latest/userguide/

API Reference: https://docs.aws.amazon.com/ram/latest/APIReference/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/ram/faqs/

Cost Optimization

AWS Resource Access Manager provides significant cost optimization by eliminating the need to duplicate resources across multiple AWS accounts. The service itself incurs no additional charges, making it a cost-effective solution for multi-account resource sharing:

- Organizations can achieve substantial cost savings by provisioning expensive resources like Transit Gateways, dedicated hosts, or licensed software once and sharing them across accounts rather than provisioning multiple instances
- This approach reduces both infrastructure costs and operational overhead associated with managing duplicate resources
- RAM enables efficient resource utilization by allowing multiple accounts to share the same underlying infrastructure while maintaining security and access control

Savings Considerations

Primary cost savings with AWS Resource Access Manager come from resource consolidation and elimination of duplicate provisioning across accounts. Organizations can significantly reduce infrastructure costs by sharing expensive resources like AWS License Manager configurations, AWS Transit Gateways, and Amazon Aurora clusters instead of creating separate instances per account:

- Operational savings include reduced management overhead, simplified compliance management through centralized resource governance, and decreased complexity in multi-account architectures
- RAM enables better resource utilization rates and helps organizations avoid over-provisioning by sharing underutilized resources across accounts
- The service supports cost allocation through resource tagging and provides visibility into resource usage across shared accounts

Service Name

AWS RTB Fabric

Service Overview

AWS RTB Fabric is a fully managed service designed for real-time bidding (RTB) advertising workloads. The service simplifies the process of connecting AdTech companies with their supply and demand partners to run high-volume, latency-sensitive RTB workloads on AWS. RTB Fabric provides consistent single-digit millisecond performance and up to 80% lower networking costs compared to standard networking costs. The service includes a dedicated, high-performance network environment for RTB workloads and partner integrations without requiring colocated, on-premises infrastructure or upfront commitments. RTB Fabric addresses the challenges of low-latency, cost-efficient infrastructure for RTB workloads that process millions of bid requests per second.

Key Use Cases

- Real-time bidding workloads: Processing millions of bid requests per second across publishers, supply-side platforms (SSPs), and demand-side platforms (DSPs)
- AdTech partner connectivity: Connecting with supply and demand partners including Amazon Ads, GumGum, Kargo, MobileFuse, Sovrn, TripleLift, Viant, and Yieldmo
- High-volume advertising transactions: Enabling reliable, high-speed exchange of OpenRTB requests and responses among multiple partners

Features

- **Dedicated High-Performance Network**: Private, high-performance network environment that provides single-digit millisecond latency for RTB workloads
- **Modules Support**: Capability to bring your own and partner applications securely into the compute environment, supporting containerized applications and foundation models
- **Cost-Optimized Networking**: Up to 80% reduction in standard cloud networking costs with pricing model aligned with RTB economics
- **Requester and Responder Gateways**: Infrastructure components that enable private, low-latency communication paths for OpenRTB traffic
- **Built-in Traffic Management**: Three built-in modules available to optimize traffic, improve bid efficiency, and increase bid response rates

Value Proposition

AWS RTB Fabric provides unique advantages for AdTech companies by eliminating the need for colocated, on-premises infrastructure while delivering consistent single-digit

millisecond performance. The service offers up to 80% cost reduction in networking expenses compared to standard cloud networking, making it economically viable for high-volume RTB operations:

- RTB Fabric's dedicated network environment ensures reliable, high-speed exchanges critical for processing millions of bid requests per second
- The service requires no upfront commitments and includes built-in modules that enhance transaction efficiency and bidding effectiveness
- Additionally, RTB Fabric simplifies connectivity to established AdTech partners and provides a pricing model specifically aligned with RTB economics, making it ideal for companies looking to optimize auction execution and maximize supply monetization

Service Integration

AWS RTB Fabric integrates with core AWS identity and access management through AWS Identity and Access Management (IAM) for permissions and security policies. The service leverages AWS's global infrastructure for multi-region deployments across US East (N:

- Virginia), US West (Oregon), Asia Pacific (Singapore), Asia Pacific (Tokyo), Europe (Frankfurt), and Europe (Ireland)
- RTB Fabric supports containerized applications, indicating integration with container orchestration services
- The service includes modules that can work with foundation models, suggesting integration with AWS AI/ML services
- Resource management is handled through Amazon Resource Names (ARNs) and supports AWS tagging for resource organization and cost allocation

Onboarding and Offboarding

Customers get started with AWS RTB Fabric by creating requester and responder gateways to establish private, low-latency communication paths for OpenRTB traffic. The onboarding process involves configuring these gateways and setting up links between RTB applications:

- Users can then leverage built-in modules to manage traffic efficiently and optimize bidding performance
- The service supports both inbound and outbound external links for connecting to various AdTech partners
- Configuration is managed through the AWS RTB Fabric API, which provides operations for creating, updating, and managing gateways and links
- The service requires proper IAM permissions and can be deployed across multiple supported AWS regions based on business requirements

Getting Started Guide: https://docs.aws.amazon.com/rtb-fabric/latest/userguide/

Data Removal

Based on the available documentation, AWS RTB Fabric provides standard AWS data deletion capabilities through its API operations. Customers can delete requester gateways, responder gateways, links, and external links using the respective API operations. The service follows AWS standard practices for resource cleanup and data removal when resources are deleted. Complete offboarding involves systematically removing all created resources including gateways and links to ensure clean teardown of the RTB infrastructure.

Backup/Restore and Disaster Recovery

AWS RTB Fabric is available across multiple AWS regions including US East (N. Virginia), US West (Oregon), Asia Pacific (Singapore), Asia Pacific (Tokyo), Europe (Frankfurt), and Europe (Ireland). This multi-region availability enables customers to implement disaster recovery strategies by deploying RTB workloads across different regions. The service leverages AWS's underlying infrastructure resilience and availability zones within each region for high availability and fault tolerance.

Backup Capabilities

AWS RTB Fabric does not appear to have specific built-in backup capabilities mentioned in the available documentation. The service is focused on providing real-time bidding infrastructure and network connectivity rather than data storage that would require traditional backup mechanisms. Configuration and gateway settings would be managed through the API and can be recreated as needed through infrastructure as code practices.

Disaster Recovery

AWS RTB Fabric supports disaster recovery through its multi-region deployment capability across six AWS regions. Customers can deploy RTB workloads in multiple regions to ensure business continuity:

- The service's architecture allows for failover scenarios by leveraging AWS's global infrastructure
- However, specific integration with AWS Elastic Disaster Recovery is not explicitly documented in the available materials

Recovery Objectives

AWS RTB Fabric helps customers meet recovery objectives through its multi-region availability and high-performance network design. The service's single-digit millisecond latency requirements and dedicated network infrastructure provide the foundation for achieving aggressive RTO targets. By deploying across multiple regions, customers can implement active-active or warm standby patterns to minimize downtime and maintain continuous RTB operations even during regional disruptions.

Service Constraints

Service Quotas: Information not available in current documentation

FAQ: Information not available in current documentation

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/rtb-fabric/latest/userguide/

User Guide: https://docs.aws.amazon.com/rtb-fabric/latest/userguide/

API Reference: https://docs.aws.amazon.com/rtb-fabric/latest/api/

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/rtb-fabric/pricing/

Cost Optimization

AWS RTB Fabric offers significant cost optimization specifically designed for RTB economics with up to 80% reduction in standard cloud networking costs. The service eliminates the need for expensive colocated, on-premises infrastructure while maintaining the low-latency requirements of RTB workloads:

- Built-in traffic management modules help optimize bid efficiency and increase response rates, potentially improving return on advertising spend
- The pricing model is aligned with RTB economics, making it cost-effective for high-volume advertising transactions
- No upfront commitments are required, allowing for flexible cost management based on actual usage patterns

Savings Considerations

The primary cost savings with AWS RTB Fabric come from eliminating traditional colocation expenses and on-premises infrastructure investments required for RTB workloads. The service's 80% networking cost reduction compared to standard cloud networking represents substantial operational savings for high-volume bidding operations:

- Built-in modules reduce the need for separate traffic management solutions, consolidating costs into a single service
- The pay-as-you-go model without upfront commitments allows for better cash flow management and cost predictability

- Multi-region deployment capabilities enable cost optimization by placing workloads in regions with favorable economics while maintaining performance requirements

Service Name

AWS Secrets Manager

Service Overview

AWS Secrets Manager is a fully managed service that helps manage, retrieve, and rotate database credentials, application credentials, OAuth tokens, API keys, and other secrets throughout their lifecycles. The service eliminates the need to hard-code credentials in application source code, reducing the risk of compromise. Secrets Manager encrypts secrets at rest using AWS KMS and secures transmission over TLS. It enables automatic rotation of secrets, replacing long-term credentials with short-term ones, and allows applications to retrieve credentials dynamically at runtime through API calls.

Key Use Cases

- Database credential management: Store and automatically rotate credentials for Amazon RDS, DocumentDB, Redshift, and other database services without disrupting applications
- Application security: Replace hard-coded API keys, OAuth tokens, and application credentials with dynamic retrieval from Secrets Manager
- Third-party SaaS integration: Manage and automatically rotate credentials for external services like Salesforce, Snowflake, and other SaaS providers
- Multi-region disaster recovery: Replicate secrets across multiple AWS regions for business continuity and low-latency access
- Compliance and auditing: Meet regulatory requirements with comprehensive logging, monitoring, and fine-grained access controls for sensitive credentials

Features

- **Automatic Rotation**: Schedule-based or on-demand rotation of secrets without disrupting applications, with built-in support for AWS services and third-party providers
- **Encryption**: End-to-end encryption at rest using AWS KMS keys and in transit using TLS, with support for customer-managed keys
- **Cross-Region Replication**: Automatic replication of secrets to multiple AWS regions for disaster recovery and cross-regional redundancy
- **Fine-Grained Access Control**: Integration with IAM for granular permissions, resource-based policies, and support for attribute-based access control
- **Managed External Secrets**: Lambda-free managed rotation for third-party SaaS providers with predefined secret formats and integrated partner ecosystem
- **Comprehensive Auditing**: Complete logging through AWS CloudTrail, monitoring via CloudWatch, and integration with security services like GuardDuty

- **Caching Support**: Built-in caching libraries and Lambda extensions to improve performance and reduce costs when retrieving secrets
- **Batch Operations**: Ability to retrieve multiple secrets in a single API call for improved efficiency in applications

Value Proposition

AWS Secrets Manager provides a comprehensive, fully managed solution for secret management that eliminates operational overhead while enhancing security. Unlike hard-coding credentials or manual rotation processes, Secrets Manager offers automatic rotation, enterprise-grade encryption, and seamless integration with AWS services:

- Key differentiators include Lambda-free managed rotation for third-party services, cross-region replication for disaster recovery, fine-grained access controls, and comprehensive auditing capabilities
- The service's pay-as-you-go pricing model with no setup fees, combined with built-in caching support and high API rate limits, makes it cost-effective for applications of any scale while meeting strict compliance requirements

Service Integration

AWS Secrets Manager integrates deeply with core AWS services to provide seamless credential management. Primary integrations include Amazon RDS, DocumentDB, and Redshift for automatic database credential rotation; AWS Lambda for secret retrieval through extensions and Powertools; Amazon EKS for Pod Identity integration; and AWS IAM for access control:

- The service works with AWS KMS for encryption key management, CloudTrail for audit logging, CloudWatch for monitoring, and GuardDuty for security threat detection
- Additional integrations span compute services like ECS and App Runner, data services like Athena and EMR, and developer tools like CodeBuild, enabling centralized secret management across the entire AWS ecosystem

Onboarding and Offboarding

Getting started with AWS Secrets Manager involves creating secrets through the AWS console, CLI, or SDKs, with built-in tutorials for common scenarios like moving hard-coded credentials and setting up database rotation. Users need the SecretsManagerReadWrite managed policy and can begin with tutorials for hardcoded secrets migration, database credential management, or rotation setup:

- The service offers multiple secret types including database credentials, API keys, and OAuth tokens, with managed external secrets available for supported partners
- Applications integrate by replacing hard-coded credentials with API calls to retrieve secrets dynamically, utilizing caching libraries for performance optimization

Getting Started Guide:
https://docs.aws.amazon.com/secretsmanager/latest/userguide/intro.html

Data Removal

Secret deletion in AWS Secrets Manager includes a configurable recovery window (7-30 days) during which secrets can be restored before permanent deletion. Customers can delete secrets immediately by setting the recovery window to zero, though this prevents restoration. The DeleteSecret API operation removes all secret versions and metadata. When offboarding, replicated secrets across regions must be deleted separately. CloudTrail logs capture all deletion activities for audit purposes.

Backup/Restore and Disaster Recovery

AWS Secrets Manager provides automatic cross-region replication for disaster recovery, maintaining synchronized copies of secrets and their metadata across multiple regions. Replicated secrets have the same ARN with region-specific endpoints and support automatic rotation propagation. The service maintains versioned secrets with labeled versions (AWSCURRENT, AWSPREVIOUS, AWSPENDING) enabling rollback capabilities. Secret replication improves workload resilience and supports low-latency access requirements across geographic regions.

Backup Capabilities

Secrets Manager implements versioning as its backup mechanism, maintaining multiple versions of each secret with staging labels. The service automatically retains previous secret versions during rotation and allows manual version management:

- Cross-region replication serves as geographic backup, with replicas maintaining independent versions while staying synchronized with the primary secret
- The service does not integrate with AWS Backup as traditional backup is replaced by versioning and replication features

Disaster Recovery

While AWS Secrets Manager doesn't directly integrate with AWS Elastic Disaster Recovery, it provides essential disaster recovery capabilities through cross-region replication and high availability design. Replicated secrets automatically synchronize across regions, enabling applications to failover seamlessly:

- The service's regional endpoints ensure continued access during regional impairments
- Multi-AZ deployment and 99.95% SLA provide resilience within regions, while cross-region replication supports broader disaster recovery strategies

Recovery Objectives

AWS Secrets Manager helps achieve low RPO and RTO objectives through cross-region replication with near real-time synchronization, eliminating data loss during regional failures. The service's high availability design with 99.95% SLA and multi-AZ deployment minimizes downtime. Cached secret retrieval and high API rate limits (10,000 requests/second for GetSecretValue) ensure rapid secret access during recovery scenarios. Automatic rotation ensures secrets remain current even during extended outages.

Service Constraints

Service Quotas:
https://docs.aws.amazon.com/secretsmanager/latest/userguide/reference_limits.html

FAQ: https://aws.amazon.com/secrets-manager/faqs/

Technical Requirements

Service Documentation:
https://docs.aws.amazon.com/secretsmanager/latest/userguide/intro.html

User Guide: https://docs.aws.amazon.com/secretsmanager/latest/userguide/intro.html

API Reference:
https://docs.aws.amazon.com/secretsmanager/latest/apireference/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/secrets-manager/pricing/

Cost Optimization

AWS Secrets Manager offers several cost optimization strategies: utilize the free tier for new customers with $200 credits; implement caching to reduce API calls and associated costs; use cost allocation tags to track and allocate secret usage across business units; consider secret consolidation where appropriate to reduce per-secret monthly charges; leverage managed rotation to eliminate Lambda function costs for supported services; and monitor unused secrets for deletion. The pay-as-you-go model charges $0.40 per secret per month plus $0.05 per 10,000 API calls, making cost tracking granular and predictable.

Savings Considerations

Primary cost savings come from eliminating manual credential management overhead and reducing security incident risks. Automated rotation eliminates operational costs of manual updates and application deployments:

- Cross-region replication reduces disaster recovery complexity compared to custom solutions
- Managed external secrets eliminate Lambda function development and maintenance costs
- Caching reduces API call costs significantly for high-frequency applications
- Cost allocation tags enable precise chargeback to business units, improving cost accountability
- The service's integration with AWS services reduces development time and operational complexity, delivering substantial indirect cost savings through improved operational efficiency

Service Name

AWS Security Hub CSPM

Service Overview

AWS Security Hub Cloud Security Posture Management (CSPM) is a comprehensive cloud security service that helps organizations manage and assess their security state in AWS by collecting security data, analyzing trends, and identifying priority security issues. The service performs automated security best practice checks against AWS resources to identify misconfigurations and evaluate security posture, with alignment to industry and regulatory frameworks including AWS Foundational Security Best Practices, CIS, PCI DSS, and NIST. Security Hub CSPM provides centralized visibility across multiple AWS accounts and regions, consolidates findings from various AWS security services and third-party products, and offers automation features for triaging and remediation through integration with Amazon EventBridge.

Key Use Cases

- Continuous Security Posture Assessment: Automatically evaluate AWS resources against security best practices and compliance frameworks with real-time monitoring and scoring
- Multi-Account Security Management: Centrally manage security across AWS Organizations with delegated administration and cross-region aggregation of findings
- Compliance Monitoring and Reporting: Track compliance status against industry standards like CIS, PCI DSS, and NIST with detailed control findings and security scores

Features

- **Automated Security Checks**: Performs continuous security best practice checks against AWS resources using prepackaged security standards and custom controls
- **Cross-Region Aggregation**: Consolidates findings and provides unified view across multiple AWS regions from a single aggregation region
- **Security Score Calculation**: Evaluates security posture with automated scoring based on control status across all enabled standards
- **Central Configuration**: Enables delegated administrators to configure Security Hub CSPM settings across multiple accounts and organizational units
- **Automation Rules**: Automatically updates findings, triggers responses, and integrates with ticketing, SOAR, and SIEM tools
- **Finding Consolidation**: Aggregates findings from AWS services and partner integrations in standardized AWS Security Finding Format (ASFF)

Value Proposition

Security Hub CSPM provides unmatched value through its comprehensive approach to cloud security posture management with automated, continuous security checks across multiple compliance frameworks simultaneously. Unlike standalone security tools, it offers centralized multi-account management through AWS Organizations integration, reducing operational overhead while improving security coverage:

- The service eliminates the complexity of managing disparate security tools by consolidating findings from multiple AWS services and third-party products into a single standardized format
- Key differentiators include real-time security scoring, cross-region aggregation capabilities, and deep integration with the AWS ecosystem for automated remediation workflows

Service Integration

Security Hub CSPM integrates deeply with core AWS services to provide comprehensive security coverage. Primary integrations include AWS Config for resource configuration monitoring and security checks, Amazon GuardDuty for threat detection findings, Amazon Inspector for vulnerability assessments, and Amazon Macie for sensitive data discovery:

- The service also integrates with AWS Organizations for multi-account management, Amazon EventBridge for automation workflows, AWS CloudTrail for audit logging, and IAM Access Analyzer for access management insights
- Additional integrations include AWS Systems Manager for patch management findings, AWS Firewall Manager for security policy compliance, and Amazon Detective for security investigation capabilities

Onboarding and Offboarding

Getting started with Security Hub CSPM involves enabling the service through the AWS Console, API, or CLI, with two primary setup methods available. Organizations can integrate with AWS Organizations by designating a delegated administrator account and enabling central configuration for multi-account management, or manually enable the service and invite member accounts individually:

- Essential prerequisites include enabling AWS Config with appropriate resource recording in each region where Security Hub CSPM will be used
- Initial setup automatically enables default security standards (AWS Foundational Security Best Practices and CIS AWS Foundations Benchmark) unless configured otherwise
- The service offers a 30-day free trial period to evaluate capabilities before charges begin

Getting Started Guide:
https://docs.aws.amazon.com/securityhub/latest/userguide/securityhub-settingup.html

Data Removal

Offboarding from Security Hub CSPM involves disabling the service through the console or API, which deactivates all standards and controls, stops finding generation and ingestion. Active findings are permanently deleted after 90 days and archived findings after 30 days. To retain findings during offboarding, customers can export them to S3 using custom actions with EventBridge rules before disabling the service. Multi-account environments require disassociating member accounts before the administrator account can be disabled.

Backup/Restore and Disaster Recovery

Security Hub CSPM does not provide traditional backup and restore capabilities as it is a security posture management service that generates findings based on current resource configurations. The service maintains finding history and audit trails, but does not store customer data that requires backup. Findings and configuration data are managed by AWS and replicated across availability zones within each region for high availability.

Backup Capabilities

Security Hub CSPM does not have built-in backup capabilities and does not integrate with AWS Backup service. The service focuses on security posture assessment rather than data backup. Finding data and configurations are managed internally by AWS with built-in redundancy and high availability, but customers cannot perform traditional backup operations on this service data.

Disaster Recovery

Security Hub CSPM does not directly support disaster recovery scenarios and does not integrate with AWS Elastic Disaster Recovery. The service is designed for security monitoring and assessment rather than business continuity. AWS manages the service infrastructure with built-in resilience, but customers should focus on disaster recovery for the underlying AWS resources being monitored by Security Hub CSPM.

Recovery Objectives

Security Hub CSPM does not directly contribute to RPO and RTO objectives as it is not a backup or disaster recovery service. However, it supports business continuity by continuously monitoring security posture and identifying vulnerabilities that could impact recovery processes. The service helps maintain security during disaster recovery scenarios by ensuring recovered resources meet security best practices and compliance requirements.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/general/latest/gr/sechub.html

FAQ: https://aws.amazon.com/security-hub/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/securityhub/latest/userguide/what-is-securityhub.html

User Guide: https://docs.aws.amazon.com/securityhub/latest/userguide/

API Reference: https://docs.aws.amazon.com/securityhub/1.0/APIReference/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/security-hub/cspm/pricing/

Cost Optimization

Security Hub CSMP offers several unique cost optimization opportunities including selective control disabling to reduce security check volumes and associated charges. Organizations can disable controls for global resources in non-primary regions to avoid duplicate charges while maintaining comprehensive coverage:

- The service provides a cost estimator tool for predicting monthly expenses across integrated security services
- Volume pricing tiers reduce per-check costs as usage scales
- Strategic use of central configuration eliminates redundant administrative overhead across multiple accounts
- Customers can optimize AWS Config recording settings specifically for Security Hub CSPM to reduce associated Config charges while maintaining security coverage

Savings Considerations

Primary cost savings stem from consolidating multiple security tools into a single platform, reducing licensing and operational costs. The 30-day free trial allows thorough evaluation before commitment, and volume pricing provides economies of scale for large deployments:

- Integration with existing AWS services eliminates need for third-party security tools and associated data transfer costs

- Automated finding correlation reduces manual security analysis time and associated labor costs
- Central configuration across AWS Organizations significantly reduces administrative overhead compared to managing security tools across individual accounts
- Organizations can strategically disable unnecessary controls and optimize AWS Config resource recording to minimize ancillary service charges while maintaining comprehensive security coverage

Service Name

AWS Serverless Application Repository

Service Overview

AWS Serverless Application Repository is a managed service that makes it easy for developers and enterprises to discover, deploy, and publish serverless applications in the AWS Cloud. The service allows users to publish applications publicly to the community or privately within teams and organizations. Applications are defined using AWS Serverless Application Model (AWS SAM) templates and can be deployed with a few clicks after configuring required fields. The repository is deeply integrated with the AWS Lambda console, enabling developers to browse applications by categories, keywords, publishers, or event sources. Users can share code samples, components, and complete applications for various serverless use cases, accelerating development and promoting reuse of proven patterns.

Key Use Cases

- Use case 1: Publishing and sharing serverless applications publicly with the developer community or privately within teams and organizations
- Use case 2: Discovering and deploying pre-built serverless applications for common patterns like web backends, data processing, chatbots, and IoT applications
- Use case 3: Creating nested applications to reuse common serverless components and patterns across multiple applications, reducing duplication and enhancing organization

Features

- **Application Publishing**: Publish serverless applications using AWS Management Console, AWS SAM CLI, or AWS SDKs with AWS SAM templates
- **Application Discovery and Deployment**: Browse and search applications by categories, keywords, publishers, or event sources with one-click deployment
- **Access Control and Sharing**: Control application visibility with private, privately shared, or public access levels including AWS Organization integration
- **Nested Applications**: Create and deploy applications that contain other applications as components for code reuse and modularity
- **Lambda Console Integration**: Deep integration with AWS Lambda console for seamless serverless application management
- **Verified Author Badge**: Author verification system to establish trust and credibility for published applications

Value Proposition

AWS Serverless Application Repository accelerates serverless development by providing a centralized marketplace for discovering, sharing, and deploying proven serverless patterns. The service eliminates the need to build common functionality from scratch, reducing development time and effort:

- With deep Lambda console integration, developers can deploy applications with minimal configuration
- The repository supports both public community sharing and private organizational use, enabling knowledge sharing while maintaining security
- Verified author badges and application validation provide trust and quality assurance
- The service supports complex architectures through nested applications, enabling modular design and component reuse across multiple projects

Service Integration

AWS Serverless Application Repository integrates primarily with AWS Lambda for serverless compute functions and AWS CloudFormation for infrastructure deployment and stack management. Applications are deployed as CloudFormation stacks, enabling unified resource management:

- The service works closely with AWS SAM (Serverless Application Model) for application definition and packaging
- Integration with Amazon API Gateway enables API-based serverless applications, while Amazon DynamoDB integration supports database requirements
- Amazon S3 provides storage for code packages and deployment artifacts
- Additional integrations include AWS IAM for access control, AWS CodePipeline for CI/CD workflows, and AWS Organizations for enterprise-level application sharing and governance

Onboarding and Offboarding

Customers begin by accessing the AWS Serverless Application Repository through the AWS Management Console or AWS Lambda console. For deploying applications, users browse or search the repository, select an application, configure required parameters, acknowledge any necessary permissions (IAM roles, resource policies), and deploy with a few clicks:

- For publishing applications, users need a valid AWS account, AWS SAM template, packaged application, source code URL, and README file
- Applications can be published using AWS Management Console, AWS SAM CLI, or AWS SDKs

- The service provides quick start tutorials and integration with GitHub for automated publishing workflows

Getting Started Guide:
https://docs.aws.amazon.com/serverlessrepo/latest/devguide/serverlessrepo-quick-start.html

Data Removal

Applications deployed from the repository are managed as AWS CloudFormation stacks, allowing complete removal through standard CloudFormation deletion processes. Publishers can delete their applications using the AWS Management Console or AWS CLI, which removes the application from the repository. For deployed applications, users delete the corresponding CloudFormation stack to remove all associated resources. Application artifacts stored in S3 are subject to standard S3 deletion processes.

Backup/Restore and Disaster Recovery

AWS Serverless Application Repository does not provide specific backup and disaster recovery capabilities as it primarily serves as a distribution mechanism for serverless applications. Published applications and their metadata are stored across multiple AWS regions for public applications. The underlying infrastructure relies on AWS's standard durability and availability guarantees. Deployed applications inherit backup capabilities from their constituent services like Lambda, DynamoDB, and S3.

Backup Capabilities

The service itself does not offer traditional backup capabilities. Published applications are replicated across regions for public availability:

- Code packages are stored in S3 with standard S3 durability
- The repository maintains application metadata and versions, but does not integrate with AWS Backup service
- Backup responsibility lies with the individual AWS services used by deployed applications

Disaster Recovery

AWS Serverless Application Repository does not directly support AWS Elastic Disaster Recovery integration. The service provides regional availability for public applications, automatically copying deployment artifacts to destination regions when applications are deployed:

- Disaster recovery depends on the underlying services used by deployed applications
- The repository itself benefits from AWS's infrastructure resilience across multiple regions

Recovery Objectives

The service contributes to recovery objectives primarily through its regional replication of public applications and rapid deployment capabilities. Applications can be quickly redeployed in different regions to support disaster recovery scenarios. However, specific RPO and RTO objectives depend on the architecture of individual serverless applications deployed from the repository rather than the repository service itself.

Service Constraints

Service Quotas:
https://docs.aws.amazon.com/serverlessrepo/latest/devguide/quotas.html

FAQ: https://aws.amazon.com/serverless/serverlessrepo/faqs/

Technical Requirements

Service Documentation:
https://docs.aws.amazon.com/serverlessrepo/latest/devguide/what-is-serverlessrepo.html

User Guide: https://docs.aws.amazon.com/serverlessrepo/latest/devguide/what-is-serverlessrepo.html

API Reference:
https://docs.aws.amazon.com/serverlessrepo/latest/devguide/resources.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/serverless/serverlessrepo/faqs/

Cost Optimization

AWS Serverless Application Repository incurs no direct charges for browsing, deploying, or publishing applications. Cost optimization focuses on the underlying AWS resources used by deployed applications:

- The service provides 5GB of free S3 storage per region for code packages
- Cost efficiency comes from reusing proven serverless patterns rather than building from scratch, reducing development time and resource consumption
- Publishers can optimize costs by sharing applications to avoid duplication across teams
- The pay-per-use nature of underlying serverless services like Lambda provides automatic cost optimization based on actual usage

Savings Considerations

Primary cost savings derive from accelerated development cycles through application reuse and proven patterns. Teams avoid duplicating development effort by leveraging existing applications from the repository:

- The serverless architecture of deployed applications provides cost efficiency through pay-per-execution billing models
- Organizations can reduce development costs by sharing applications internally, eliminating redundant development across teams
- The service's integration with serverless services enables automatic scaling and cost optimization based on demand
- No infrastructure management overhead reduces operational costs compared to traditional application deployment models

Service Name

AWS Shield

Service Overview

AWS Shield is a managed Distributed Denial of Service (DDoS) protection service that safeguards web applications running on AWS. It provides two tiers: Shield Standard (free) and Shield Advanced (paid subscription). Shield Standard automatically protects against common network and transport layer (Layer 3/4) DDoS attacks for services like CloudFront, Route 53, and Elastic Load Balancing. Shield Advanced offers enhanced protection with application layer (Layer 7) DDoS mitigation, 24/7 access to the AWS DDoS Response Team, advanced event visibility, proactive engagement, and cost protection against DDoS-induced usage spikes.

Key Use Cases

- Web application protection: Protecting internet-facing web applications from DDoS attacks that could cause downtime or service degradation
- E-commerce and critical business applications: Ensuring high availability for revenue-generating applications that require guaranteed uptime during attack scenarios
- Gaming and media streaming: Protecting applications with unpredictable traffic patterns from volumetric attacks and ensuring consistent user experience
- API and microservices protection: Safeguarding API gateways and distributed architectures from application layer request floods and sophisticated attacks

Features

- **Shield Standard**: Free automatic protection against common Layer 3/4 DDoS attacks with static threshold detection and inline mitigation
- **Shield Advanced**: Enhanced DDoS protection with application layer mitigation, health-based detection, and tailored traffic pattern analysis
- **AWS WAF Integration**: Automatic application layer DDoS mitigation using AWS WAF rules and rate-based controls included at no additional cost
- **Shield Response Team (SRT)**: 24/7 access to AWS DDoS experts for attack response, custom mitigation creation, and proactive engagement during incidents
- **DDoS Cost Protection**: Service credits to cover scaling charges resulting from DDoS attacks against protected resources
- **Advanced Event Visibility**: Real-time metrics, detailed attack reports, and CloudWatch integration for comprehensive DDoS event monitoring
- **Protection Groups**: Logical groupings of protected resources for enhanced detection, mitigation, and centralized management

- **Network Security Director**: Preview capability providing network topology visibility, security configuration analysis, and remediation recommendations

Value Proposition

AWS Shield provides comprehensive DDoS protection with the advantage of being deeply integrated into AWS infrastructure and services. Shield Standard is automatically included at no cost, providing immediate protection without configuration:

- Shield Advanced offers unique value through its AWS DDoS Response Team access, cost protection guarantees, and integration with AWS WAF at no additional charge for standard capabilities
- The service leverages AWS's massive global infrastructure scale for mitigation capacity and provides proactive engagement during attacks
- Unlike third-party solutions, Shield offers seamless integration with other AWS services, centralized management through AWS Firewall Manager, and the ability to protect both AWS-hosted and external resources through CloudFront and Route 53

Service Integration

AWS Shield integrates tightly with multiple core AWS services for comprehensive protection. Amazon CloudFront distributions and Amazon Route 53 hosted zones receive enhanced protection automatically:

- AWS WAF integration provides application layer mitigation included with Shield Advanced subscriptions
- Elastic Load Balancing (Application, Network, and Classic Load Balancers) can be protected when associated with Elastic IP addresses
- Amazon EC2 instances are protected through Elastic IP association
- AWS Global Accelerator standard accelerators receive automatic protection
- AWS Firewall Manager enables centralized Shield Advanced policy management across accounts
- CloudWatch provides metrics and monitoring for protected resources
- AWS Organizations supports consolidated billing for Shield Advanced across multiple accounts

Onboarding and Offboarding

Getting started with AWS Shield requires different approaches for each tier. Shield Standard is automatically enabled for all AWS customers at no cost and requires no configuration:

- For Shield Advanced, customers must subscribe through the AWS console or API, requiring a one-year commitment with $3,000 monthly fee plus usage charges

- The setup process involves subscribing to Shield Advanced, adding resource protections, configuring health-based detection with Route 53 health checks, setting up CloudWatch alarms and notifications, and optionally configuring AWS Shield Response Team support with emergency contacts
- Shield Advanced can be deployed using Infrastructure as Code tools including AWS CloudFormation and Terraform for automated configuration

Getting Started Guide: https://docs.aws.amazon.com/waf/latest/developerguide/getting-started-ddos.html

Data Removal

AWS Shield offboarding varies by tier. Shield Standard cannot be disabled as it's automatically included with AWS services. For Shield Advanced, customers can cancel subscriptions through the AWS console, but the one-year commitment must be fulfilled. During offboarding, all Shield Advanced protections are removed, custom mitigations are deleted, and SRT access is terminated. No customer data is stored by Shield itself, as it only analyzes traffic patterns in real-time. Historical attack and event data in CloudWatch follows standard CloudWatch retention policies and can be deleted according to customer preferences.

Backup/Restore and Disaster Recovery

AWS Shield does not provide backup and disaster recovery capabilities as it is a real-time DDoS protection service. Shield operates at the network perimeter and does not store or backup customer data. The service itself is designed for high availability across AWS's global infrastructure with automatic failover capabilities. For disaster recovery planning, customers should implement Shield protections in multiple regions and combine with other AWS services like AWS Backup for data protection and AWS Elastic Disaster Recovery for application recovery scenarios.

Backup Capabilities

AWS Shield does not have backup capabilities and does not integrate with AWS Backup. Shield is a real-time traffic analysis and mitigation service that does not store customer data requiring backup:

- The service configuration and protection settings are maintained in AWS's control plane but individual attack data is transient
- Customers should rely on other AWS services for backup needs while using Shield purely for DDoS protection of their applications and infrastructure

Disaster Recovery

AWS Shield does not directly support disaster recovery scenarios and does not integrate with AWS Elastic Disaster Recovery. However, Shield provides crucial availability

protection during disasters by maintaining DDoS mitigation capabilities across AWS regions:

- In disaster recovery architectures, Shield protections should be configured for both primary and secondary regions
- Shield's high availability design ensures continued DDoS protection even during regional failures, supporting overall disaster recovery objectives by maintaining application availability during recovery operations

Recovery Objectives

AWS Shield helps customers meet aggressive RTO objectives by providing immediate, automatic DDoS mitigation without manual intervention. Shield Standard offers real-time protection with no recovery delay. Shield Advanced provides faster detection and mitigation through health-based detection and proactive SRT engagement. The service maintains continuous protection availability across AWS's global infrastructure, supporting near-zero RTO for DDoS-related incidents. For RPO objectives, Shield's real-time nature means no data loss during DDoS attacks, though customers must implement separate backup strategies for comprehensive data protection.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/general/latest/gr/shield.html

FAQ: https://aws.amazon.com/shield/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/waf/latest/developerguide/shield-chapter.html

User Guide: https://docs.aws.amazon.com/waf/latest/developerguide/getting-started-ddos.html

API Reference: https://docs.aws.amazon.com/waf/latest/DDOSAPIReference/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/shield/pricing/

Cost Optimization

AWS Shield offers several unique cost optimization features. Shield Standard is completely free for all AWS customers with no usage charges:

- Shield Advanced includes DDoS cost protection through service credits that cover usage spikes caused by DDoS attacks, preventing unexpected bills
- The subscription covers standard AWS WAF capabilities for protected resources, eliminating separate WAF charges
- Consolidated billing allows one Shield Advanced subscription to cover all accounts in an AWS Organizations family
- For multiple organizations, customers can request fee waivers to pay one subscription across all organizations
- The service's automatic mitigation reduces the need for expensive manual intervention and additional capacity provisioning during attacks

Savings Considerations

Main cost savings with AWS Shield come from avoided downtime and the DDoS cost protection feature. Shield Standard eliminates the need for third-party DDoS protection services for basic protection:

- Shield Advanced prevents revenue loss from DDoS-induced outages while providing cost protection against scaling charges during attacks
- The included AWS WAF capabilities reduce separate security tool costs
- Automated mitigation reduces operational overhead and the need for dedicated security personnel during incidents
- The service's integration with AWS infrastructure eliminates data transfer costs between separate security appliances
- For high-value applications, the cost of Shield Advanced is typically offset by preventing a single successful DDoS attack that could cause significant business impact and recovery costs

Service Name

AWS Step Functions

Service Overview

AWS Step Functions is a serverless workflow orchestration service that enables users to create and manage complex application workflows, also known as state machines. It coordinates multiple AWS services into serverless workflows by breaking down applications into multiple steps, adding flow logic, and tracking inputs and outputs between steps. Step Functions uses Amazon States Language (ASL) to define workflows as a series of event-driven steps, where each step is a state. The service offers two workflow types: Standard workflows for long-running, auditable workflows that can run up to one year with exactly-once execution, and Express workflows for high-event-rate workloads that can run up to five minutes with at-least-once execution. Step Functions maintains application state, provides built-in error handling, supports parallel processing, and integrates with over 220 AWS services and 10,000+ APIs.

Key Use Cases

- Microservice orchestration: Combining Lambda functions to build web-based applications with or without human approval steps
- Security and IT automation: Orchestrating security incident response workflows and automating responses to operational events in AWS accounts
- Data processing and ETL orchestration: Building data processing pipelines for streaming data and automating ETL processes with parallel job execution
- Machine learning operations: Running ETL jobs and building, training, and deploying machine learning models using Amazon SageMaker integration
- Media processing: Extracting data from PDFs or images and building serverless video transcoding pipelines with massive parallelization

Features

- **Visual Workflow Studio**: Drag-and-drop interface with Design, Code, and Config modes for creating and editing workflows visually or through Amazon States Language
- **Two Workflow Types**: Standard workflows for long-running auditable processes (up to 1 year) and Express workflows for high-volume short-running processes (up to 5 minutes)
- **Service Integration Patterns**: Three integration patterns: Request Response, Run a Job (.sync), and Wait for Callback (.waitForTaskToken) supporting 220+ AWS services and 10,000+ APIs

- **Built-in Error Handling**: Automatic retry mechanisms, dead-letter queues, error notifications, and exception handling with Catch and Retry states
- **State Management**: Maintains workflow state with execution history, input/output tracking, and ability to resume after interruptions
- **Parallel Processing**: Map state in Distributed mode for large-scale parallel processing and Parallel state for concurrent execution branches
- **Monitoring and Observability**: Integration with CloudWatch for logging, X-Ray for tracing, and visual execution history with detailed event tracking

Value Proposition

AWS Step Functions provides a serverless, visual approach to orchestrating complex workflows without managing infrastructure. Unlike traditional workflow engines or custom orchestration code, Step Functions eliminates the need to write connection logic between services, provides built-in error handling and retry mechanisms, and offers pay-per-use pricing:

- The service accelerates application development through its intuitive visual interface and pre-built integrations with 220+ AWS services
- Step Functions ensures high availability and fault tolerance with automatic scaling, maintains detailed execution history for auditing and debugging, and supports both synchronous and asynchronous execution patterns
- The service simplifies modernization of distributed applications by providing a centralized orchestration layer that reduces complexity and improves maintainability while offering enterprise-grade security and compliance features

Service Integration

Step Functions integrates seamlessly with core AWS compute services including Lambda for serverless functions, ECS and Fargate for containerized workloads, and AWS Batch for batch processing jobs. Data services integration includes DynamoDB for NoSQL operations, S3 for object storage, RDS for relational databases, and Amazon Redshift for data warehousing:

- Machine learning integrations cover Amazon SageMaker for model training and inference, Amazon Bedrock for generative AI, and Amazon Comprehend for natural language processing
- Messaging and event services include SNS for notifications, SQS for queuing, EventBridge for event routing, and API Gateway for REST APIs
- Additional integrations span monitoring services like CloudWatch, security services like IAM, and developer tools like CodeBuild and CodePipeline, providing comprehensive orchestration capabilities across the AWS ecosystem

Onboarding and Offboarding

Customers can get started with Step Functions through the AWS Console's built-in Hello World workflow tutorial, which takes 20-30 minutes to complete. The process begins by accessing the Step Functions console and selecting 'Get Started' followed by 'Run Hello World' to create a simple workflow using Workflow Studio's drag-and-drop interface:

- The tutorial covers creating state machines using visual design mode, understanding flow states and task states, viewing Amazon States Language (ASL) code, and integrating AWS services like Amazon Comprehend
- Customers learn to configure IAM roles and policies, execute workflows with input data, and monitor execution details
- The onboarding includes hands-on experience with the three Workflow Studio modes: Design for visual editing, Code for direct ASL editing, and Config for state machine configuration including logging, tracing, and security settings

Getting Started Guide: https://docs.aws.amazon.com/step-functions/latest/dg/getting-started.html

Data Removal

Step Functions data removal involves deleting state machine definitions, execution history, and associated resources. Customers can delete state machines through the console, CLI, or API, which removes the workflow definition and stops new executions. Existing execution history is retained according to service quotas but can be managed through CloudWatch Logs retention policies if logging is enabled. For complete offboarding, customers must also remove associated IAM roles, CloudWatch log groups, and any Lambda functions or other resources created specifically for the workflows to ensure no residual charges or data retention.

Backup/Restore and Disaster Recovery

Step Functions provides inherent resilience through its serverless architecture with automatic multi-AZ deployment and built-in fault tolerance. State machine definitions are automatically backed up as part of the AWS control plane. Execution state is maintained during service interruptions, allowing workflows to resume from the last successful state. For disaster recovery, customers can export state machine definitions as JSON/YAML and redeploy in alternate regions. Cross-region replication of dependent resources like Lambda functions and DynamoDB tables ensures complete DR coverage, while execution history can be preserved through CloudWatch Logs cross-region replication.

Backup Capabilities

Step Functions automatically maintains state machine definitions within the AWS control plane infrastructure. The service does not directly integrate with AWS Backup as it's a serverless orchestration service rather than a data storage service:

- However, customers can programmatically export state machine definitions using APIs for backup purposes
- The service maintains execution history up to quota limits, and when CloudWatch logging is enabled, execution logs are preserved according to CloudWatch retention policies, providing audit trails and debugging information for recovery scenarios

Disaster Recovery

Step Functions supports disaster recovery through its multi-region availability and serverless architecture. The service does not directly integrate with AWS Elastic Disaster Recovery as it's an orchestration service rather than compute infrastructure:

- However, customers can implement DR by maintaining state machine definitions in multiple regions and replicating dependent resources
- The service's stateless nature and JSON-based definitions enable quick recovery by redeploying workflows in alternate regions, while execution state resilience ensures workflows can continue after regional failures

Recovery Objectives

Step Functions helps customers achieve aggressive RPO and RTO objectives through its serverless architecture and built-in resilience. For RPO, the service maintains real-time state information and can resume workflows from the last successful state, minimizing data loss. For RTO, the multi-AZ deployment provides automatic failover capabilities within regions, while cross-region deployment of state machines enables rapid recovery during regional outages. The service's integration with other AWS services allows customers to design workflows that incorporate backup and recovery patterns, such as automated failover procedures and data replication workflows.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/general/latest/gr/step-functions.html

FAQ: https://aws.amazon.com/step-functions/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/step-functions/latest/dg/welcome.html

User Guide: https://docs.aws.amazon.com/step-functions/latest/dg/getting-started.html

API Reference: https://docs.aws.amazon.com/step-functions/latest/apireference/index.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/step-functions/pricing/

Cost Optimization

Step Functions offers unique cost optimization through workflow type selection, where Express workflows cost significantly less than Standard workflows for high-volume, short-duration use cases. Customers can nest Express workflows within Standard workflows to optimize costs for mixed workloads:

- The service supports cost optimization through efficient state design, minimizing state transitions which are the primary billing unit for Standard workflows
- Using S3 ARNs instead of passing large payloads reduces state transition costs and avoids payload size limits
- Map state in Distributed mode provides cost-effective parallel processing for large-scale operations
- Proper timeout configuration prevents runaway executions and unnecessary charges, while workflow consolidation reduces overall state transitions and associated costs

Savings Considerations

Primary cost savings come from choosing appropriate workflow types, with Express workflows offering up to 90% cost reduction for high-volume, short-duration scenarios compared to Standard workflows. Optimizing state machine design by reducing unnecessary state transitions, combining multiple operations into single states, and using efficient data passing patterns significantly reduces costs:

- Leveraging the free tier of 4,000 state transitions per month for Standard workflows provides cost savings for small to medium workloads
- Using Step Functions instead of managing custom orchestration infrastructure eliminates server costs, reduces operational overhead, and provides pay-per-use billing
- The service's integration capabilities reduce development time and maintenance costs compared to building custom workflow solutions, while automatic scaling eliminates over-provisioning costs typical in traditional workflow engines

Service Name

AWS Storage Gateway

Service Overview

AWS Storage Gateway is a hybrid cloud storage service that connects on-premises software appliances with cloud-based storage to provide seamless integration between on-premises IT environments and AWS storage infrastructure. The service enables organizations to store data in the AWS Cloud for scalable and cost-effective storage while maintaining data security features. Storage Gateway can be deployed as a VM appliance on VMware ESXi, KVM, or Microsoft Hyper-V hypervisors, as a hardware appliance on-premises, or as an Amazon EC2 instance in AWS. It provides file-based, volume-based, and tape-based storage solutions using standard storage protocols such as iSCSI, SMB, and NFS, allowing organizations to use AWS storage without rewriting existing applications.

Key Use Cases

- Backup and disaster recovery: Move backups and archives to the cloud while maintaining on-premises access for quick recovery
- Hybrid cloud file shares: Reduce on-premises storage footprint with cloud-backed file shares using SMB and NFS protocols
- Data lake ingestion: Store file data as objects in Amazon S3 for data lakes, analytics, and machine learning workflows
- Application data migration: Migrate application data to AWS while providing low-latency access to frequently used data
- Virtual tape library replacement: Replace physical tape infrastructure with virtual tapes stored in Amazon S3 and Glacier storage classes

Features

- **Amazon S3 File Gateway**: Enables file storage as objects in Amazon S3 accessible via NFS and SMB protocols with local caching
- **Amazon FSx File Gateway**: Provides low-latency access to Amazon FSx for Windows File Server file shares on-premises
- **Volume Gateway (Cached Volumes)**: Stores data in Amazon S3 with frequently accessed data cached locally for low-latency access
- **Volume Gateway (Stored Volumes)**: Stores primary data locally with asynchronous backup to Amazon S3 as EBS snapshots
- **Tape Gateway**: Virtual tape library (VTL) interface for backup applications with cloud-based virtual tape storage
- **Local Caching**: Intelligent caching of frequently accessed data on-premises for low-latency performance

- **Data Optimization**: Compresses data and transfers only changed data to optimize bandwidth usage
- **Multiple Deployment Options**: Deploy as VM, hardware appliance, or EC2 instance with support for multiple hypervisors

Value Proposition

AWS Storage Gateway provides unique value by enabling seamless hybrid cloud storage without requiring application rewrites or changes to existing workflows. It reduces on-premises storage costs and complexity while maintaining low-latency access to frequently used data through intelligent local caching:

- The service offers flexible deployment options across VM, hardware appliance, and cloud-based instances, making it adaptable to various infrastructure requirements
- Storage Gateway integrates natively with AWS services like S3, CloudWatch, IAM, and KMS, providing enterprise-grade security, monitoring, and management capabilities
- Unlike pure cloud storage solutions, it bridges on-premises and cloud environments with standard protocols (iSCSI, SMB, NFS) that existing applications already understand

Service Integration

AWS Storage Gateway integrates deeply with multiple core AWS services to provide comprehensive hybrid storage capabilities. It stores data in Amazon S3 for file gateways and uses S3 storage classes including Glacier and Deep Archive for cost optimization:

- Volume Gateway creates EBS snapshots stored in S3 for backup and recovery
- The service leverages AWS IAM for access management and security, AWS KMS for encryption at rest, and integrates with AWS Backup for centralized backup management across AWS resources
- CloudWatch provides monitoring and metrics, while CloudTrail logs all API activities
- Storage Gateway also works with AWS Direct Connect for dedicated network connectivity and supports VPC endpoints for private connectivity to AWS services

Onboarding and Offboarding

Customers begin by signing up for an AWS account and creating an IAM user with administrative permissions to access Storage Gateway through the console or programmatically via AWS SDKs. The setup process involves choosing the gateway type (File, Volume, or Tape), selecting the host platform (VM, hardware appliance, or EC2), and downloading/deploying the gateway software:

- After deployment, customers activate the gateway by specifying service endpoints and IP addresses, then allocate local storage for caching and configure cloud storage resources like volumes, file shares, or virtual tapes
- The service provides setup wizards and documentation to guide users through requirements, network configuration, and initial resource creation to begin transferring data between on-premises and AWS storage

Getting Started Guide:
https://docs.aws.amazon.com/storagegateway/latest/vgw/GettingStarted.html

Data Removal

To offboard from AWS Storage Gateway, customers must first ensure no applications are writing to gateway resources to prevent data loss. The process involves deleting gateway-specific resources (volumes, file shares, or virtual tapes), followed by deleting the gateway itself through the Storage Gateway console. For on-premises deployments, customers must also remove the gateway VM from their virtualization environment. Associated AWS resources like EBS snapshots, S3 objects, and EC2 instances persist after gateway deletion and must be manually removed to avoid ongoing charges. Customers should backup critical data before deletion as gateway resources cannot be recovered once deleted.

Backup/Restore and Disaster Recovery

AWS Storage Gateway provides comprehensive backup and disaster recovery capabilities through integration with AWS Backup and native snapshot functionality. Volume Gateway creates point-in-time EBS snapshots stored in Amazon S3 for data protection and recovery. Tape Gateway offers virtual tape archives in S3 Glacier and Deep Archive for long-term backup retention. The service enables cross-region replication for disaster recovery scenarios and supports recovery to both on-premises environments and Amazon EC2 instances. All backup data benefits from S3's 99.999999999% durability and can be managed through centralized policies using AWS Backup.

Backup Capabilities

Storage Gateway provides native backup capabilities through automated EBS snapshots for Volume Gateway and integrates with AWS Backup for centralized backup management across all gateway types. The service supports both scheduled and on-demand snapshots, with incremental backups to reduce storage costs and transfer time:

- Backup data is stored in Amazon S3 with high durability and supports cross-region backup for disaster recovery
- Customers can restore volumes from snapshots to any compatible gateway or directly to EBS volumes in EC2 for cloud-based recovery scenarios

Disaster Recovery

Storage Gateway supports comprehensive disaster recovery through multiple mechanisms including cross-region snapshot replication, cloud-based recovery to EC2 instances, and integration with AWS Elastic Disaster Recovery. The service enables rapid recovery from backup snapshots stored in S3, supports failover to alternate gateway instances, and provides high availability configurations on VMware environments. Volume Gateway cached volumes can be quickly provisioned from snapshots for disaster recovery scenarios, while stored volumes provide local data protection with cloud-based backup for comprehensive DR coverage.

Recovery Objectives

AWS Storage Gateway helps customers meet RPO objectives through frequent incremental snapshots (as often as every hour) and continuous data replication to AWS storage services. RTO objectives are supported through local caching for immediate access to frequently used data, rapid snapshot restoration capabilities, and the ability to recover volumes directly to EC2 instances in the cloud. Cached volumes provide near-instantaneous access to cached data while stored volumes offer immediate access to all local data with cloud-based recovery options for comprehensive RPO/RTO coverage.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/general/latest/gr/sg.html

FAQ: https://aws.amazon.com/storagegateway/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/storagegateway/

User Guide:
https://docs.aws.amazon.com/storagegateway/latest/vgw/WhatIsStorageGateway.html

API Reference:
https://docs.aws.amazon.com/storagegateway/latest/APIReference/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/storagegateway/pricing/

Cost Optimization

AWS Storage Gateway offers several unique cost optimization features including intelligent data tiering that automatically moves infrequently accessed data to lower-cost S3 storage

classes like Glacier and Deep Archive. The service reduces data transfer costs through local caching, compression, and incremental data synchronization that only transfers changed data:

- Amazon S3 File Gateway provides up to 98% cost savings compared to traditional on-premises storage by leveraging S3's various storage classes
- FSx File Gateway can save approximately $180 per month for multi-region workloads through intelligent caching and reduced data transfer charges
- The service eliminates the need for expensive on-premises storage infrastructure while providing pay-as-you-use pricing for actual data stored and transferred

Savings Considerations

Primary cost savings come from eliminating expensive on-premises storage hardware, maintenance contracts, and infrastructure scaling costs. Storage Gateway reduces operational expenses through automated data management, eliminating the need for manual tape handling and offsite storage logistics:

- The service provides significant savings through S3's tiered storage pricing, with automatic lifecycle transitions to cheaper storage classes for infrequently accessed data
- Data compression and deduplication reduce both storage costs and bandwidth charges
- Organizations can optimize costs by right-sizing local cache storage based on access patterns and leveraging AWS Free Tier benefits
- The hybrid model allows customers to maintain local performance while benefiting from cloud economics and virtually unlimited scale without upfront capital investments

Service Name

AWS Supply Chain

Service Overview

AWS Supply Chain is a cloud-based supply chain management application that works with existing enterprise resource planning (ERP) and supply chain management systems. It connects and extracts inventory, supply, and demand data from existing systems into a unified AWS Supply Chain data model. The service uses machine learning and generative AI to enhance demand forecasting, inventory visibility, and supply planning. It provides end-to-end visibility into supply chains, enabling organizations to collaborate with suppliers, automate order processing, gain insights through analytics, and digitally transform their supply chain operations without requiring replatforming, upfront licensing fees, or long-term commitments.

Key Use Cases

- Demand Planning: Create demand forecasts, adjust forecasts according to market conditions, and enable demand planners to collaborate across teams using machine learning algorithms
- Supply Planning: Plan and forecast purchases of raw materials, components, and finished goods with auto replenishment and manufacturing plans
- Supply Chain Risk Management: Automatically generate insights of potential supply chain risks like stockouts, excess stocks, and lead time deviations with rebalance recommendations
- N-Tier Visibility: Extend visibility and insights beyond the organization to external trading partners for comprehensive supply chain transparency
- Manufacturing Operations: Transform manufacturing processes by developing effective material resource plans, monitoring material composition changes, and maintaining spare parts visibility
- Sustainability Management: Invite partners and map their sustainability information to create a secure and efficient way for obtaining documents and datasets from supplier networks

Features

- **Data Lake**: Simplifies aggregating data from supply chain systems using an extensible data model built for supply chain management, supporting Amazon S3, EDI, SAP S/4HANA, and SAP ECC 6.0
- **Insights**: Automatically generates insights of potential supply chain risks and provides rebalance recommendations using machine learning

- **Order Planning and Tracking**: View work order status, expected time of arrival predictions, delivery risk and recommendations for each work order
- **Demand Planning**: Create demand forecasts using machine learning, adjust forecasts according to market conditions, and enable collaboration across teams
- **Supply Planning**: Supports auto replenishment and manufacturing plans for planning and forecasting purchases of raw materials, components, and finished goods
- **N-Tier Visibility**: Extends visibility and insights beyond the organization to external trading partners
- **Sustainability**: Enables inviting partners and mapping their sustainability information for comprehensive environmental tracking
- **Amazon Q Integration**: Generative AI assistant integrated into AWS Supply Chain for querying and analyzing data using natural language

Value Proposition

AWS Supply Chain offers a comprehensive cloud-based solution that eliminates the need for replatforming, upfront licensing fees, or long-term commitments. It leverages AWS's secure infrastructure, machine learning, and generative AI to provide end-to-end supply chain visibility and automation:

- The service integrates seamlessly with existing ERP systems while offering advanced features like automated risk detection, demand forecasting, and supplier collaboration
- Customers benefit from reduced time-to-market, improved operational efficiency, cost optimization through pay-as-you-go pricing, and the ability to scale globally using AWS's infrastructure
- The integration with Amazon Q provides natural language querying capabilities, making complex supply chain data analysis accessible to all stakeholders

Service Integration

AWS Supply Chain integrates deeply with multiple AWS services to provide comprehensive functionality. Core integrations include Amazon S3 for data storage and retrieval of product catalogs, purchase orders, and invoices; Amazon SageMaker for building machine learning models for demand forecasting and anomaly detection; AWS Lambda for serverless code execution and event processing; Amazon DynamoDB for storing inventory levels and order status data; Amazon SNS and Amazon SQS for notifications and messaging; AWS Step Functions for coordinating workflows; Amazon Kinesis for real-time data stream processing; AWS Glue for ETL operations; AWS Lake Formation for data lake management; IAM Identity Center for authentication; and Amazon QuickSight for embedded analytics. The service also leverages AWS CloudTrail, AWS Config, and AWS Trusted Advisor for monitoring, compliance, and optimization.

Onboarding and Offboarding

Getting started with AWS Supply Chain involves a four-step process. First, assign an IAM Identity Center User profile to manage identity synchronization:

- Second, create an AWS Supply Chain instance through the AWS console
- Third, choose an AWS Supply Chain application owner to manage the instance
- Fourth, log into the AWS Supply Chain web application using provided credentials
- The service supports up to 10 instances per AWS account and offers a test drive feature with pre-populated datasets for evaluation
- Users need a broadband internet connection and supported web browsers (Chrome, Firefox, Edge, Safari)
- The platform provides various permission roles including Administrator, Data Analyst, Inventory Manager, Planner, Partner Data Manager, and Supply Planner to manage access appropriately

Getting Started Guide: https://docs.aws.amazon.com/aws-supply-chain/latest/adminguide/getting-started.html

Data Removal

When deleting an AWS Supply Chain instance, the service provides APIs for programmatic deletion including DeleteInstance, DeleteDataLakeDataset, DeleteDataIntegrationFlow, and DeleteDataLakeNamespace. Instance deletion is an asynchronous operation that cleans up AWS resources created during instance creation and deletes associated KMS keys and relevant information. However, data stored in Amazon S3 buckets is not automatically deleted and must be managed separately. Related information in Amazon Q is also deleted when the instance is removed. Users must manually handle S3 data cleanup as part of the offboarding process.

Backup/Restore and Disaster Recovery

AWS Supply Chain leverages AWS's comprehensive backup and disaster recovery infrastructure. Data stored in the service benefits from Amazon S3's durability and can be protected using AWS Backup for centralized backup management. The service supports cross-region replication capabilities inherent in underlying AWS services. However, specific backup and disaster recovery features are primarily dependent on the underlying AWS infrastructure rather than native Supply Chain-specific backup capabilities. Users should implement backup strategies using AWS services like S3 replication, DynamoDB backups, and AWS Backup for comprehensive data protection.

Backup Capabilities

AWS Supply Chain does not provide native backup capabilities but relies on the backup features of underlying AWS services. Data stored in Amazon S3 benefits from built-in durability and versioning capabilities:

- Users can leverage AWS Backup to create centralized backup policies for supported AWS services used by Supply Chain
- The service integrates with services like DynamoDB and S3 that have their own backup mechanisms
- For comprehensive data protection, customers should implement backup strategies using AWS's ecosystem of backup services rather than relying on Supply Chain-specific backup features

Disaster Recovery

AWS Supply Chain supports disaster recovery through AWS's infrastructure rather than native disaster recovery features. The service can benefit from AWS Elastic Disaster Recovery for continuous replication and AWS's multi-region architecture for resilience:

- Cross-region deployment capabilities allow for disaster recovery planning using AWS CloudFormation for infrastructure deployment
- However, the service does not integrate directly with AWS Elastic Disaster Recovery as a managed feature
- Disaster recovery planning requires implementing solutions using AWS's broader disaster recovery services and multi-region deployment strategies

Recovery Objectives

AWS Supply Chain helps customers meet RPO and RTO objectives through AWS's underlying infrastructure capabilities rather than service-specific features. Customers can achieve low RPO through continuous data replication using S3 cross-region replication, DynamoDB global tables, and other AWS service capabilities. RTO can be improved through multi-region deployments and automated failover mechanisms using services like Route 53 and AWS Application Recovery Controller. However, specific RPO and RTO targets depend on customer implementation of AWS disaster recovery best practices rather than built-in Supply Chain recovery features.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/aws-supply-chain/latest/adminguide/quotas.html

FAQ: https://docs.aws.amazon.com/aws-supply-chain/latest/adminguide/faq.html

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/aws-supply-chain/latest/userguide/index.html

User Guide: https://docs.aws.amazon.com/aws-supply-chain/latest/userguide/what-is-service.html

API Reference: https://docs.aws.amazon.com/aws-supply-chain/latest/APIReference/index.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/aws-supply-chain/pricing/

Cost Optimization

AWS Supply Chain offers several cost optimization options through its pay-as-you-go pricing model. The service provides a 30-day free trial including 1,000 product SKU location combinations for Insights and Demand Planning, 1,000 component SKU location combinations for Supply Planning, 25 data requests for Sustainability, and 10 GB of data storage:

- Cost optimization is achieved through tiered pricing that reduces per-unit costs at higher volumes, particularly for Supply Planning which ranges from $0.30 to $0.10 per combination for the first 1,000,000 combinations
- Data lake storage charges $0.25 per GB per month for additional storage beyond the free tier
- N-Tier Visibility is included in Supply Planning and Insights pricing, providing additional value without separate charges
- The service eliminates upfront licensing fees and long-term commitments typical of traditional supply chain solutions

Savings Considerations

Main cost savings considerations include eliminating traditional on-premises supply chain software licensing costs, reducing infrastructure maintenance expenses, and leveraging AWS's economies of scale. The service's cloud-native architecture eliminates the need for dedicated hardware, software maintenance, and IT staff for system management:

- Volume-based pricing tiers provide significant savings for large-scale implementations, with per-unit costs decreasing substantially at higher usage levels

- Integration with existing AWS services can reduce overall cloud costs through consolidated billing and shared resource utilization
- The pay-as-you-go model ensures customers only pay for actual usage rather than peak capacity provisioning
- Automated insights and machine learning capabilities can reduce operational costs by identifying inefficiencies and optimizing inventory levels, potentially saving significant working capital costs

Service Name

AWS Systems Manager

Service Overview

AWS Systems Manager is a comprehensive management solution that helps organizations centrally view, manage, and operate nodes at scale in AWS, on-premises, and multicloud environments. It provides a unified console experience that consolidates various tools to help complete common node tasks across AWS accounts and regions. Systems Manager offers secure end-to-end operations management, acting as the central hub for AWS applications and hybrid cloud infrastructure. The service requires nodes to be managed through the SSM Agent, which enables communication with the Systems Manager service for operational tasks like patching, automation, session management, and configuration compliance.

Key Use Cases

- Node Management: Centrally manage EC2 instances, on-premises servers, and multicloud environments through a unified console with automated patching, inventory management, and compliance monitoring
- Secure Access and Automation: Enable secure shell access without SSH/RDP, automate common administrative tasks, and execute commands remotely across multiple nodes without opening inbound ports
- Configuration Management: Store and manage application configuration data, secrets, and parameters centrally while automating software deployments and system configurations at scale

Features

- **Run Command**: Execute commands remotely on managed nodes without SSH access, with rate control and targeting capabilities
- **Session Manager**: Secure browser-based shell access to managed nodes without opening inbound ports or managing bastion hosts
- **Patch Manager**: Automate the patching process for managed nodes with security updates and other patches using patch baselines
- **Parameter Store**: Secure hierarchical storage for configuration data, secrets, and parameters with encryption and access control
- **Automation**: Automate common administrative and operational tasks using predefined or custom runbooks
- **State Manager**: Maintain consistent configuration across managed nodes by defining and enforcing desired state

- **Inventory**: Collect metadata and software inventory information from managed nodes for compliance and asset management
- **OpsCenter**: Central location for operations engineers to view, investigate, and resolve operational work items
- **Fleet Manager**: Visual interface for managing and troubleshooting servers with file system browsing and performance monitoring
- **Compliance**: Monitor and report on configuration compliance across managed nodes with integration to AWS Config

Value Proposition

AWS Systems Manager provides unparalleled operational efficiency by eliminating the need for bastion hosts, SSH access, or manual server management while offering comprehensive visibility across hybrid and multicloud environments. The service uniquely combines security, automation, and scale with its agent-based approach that requires no inbound ports, reducing attack surface while enabling fine-grained control:

- Key differentiators include native integration with AWS services, support for on-premises and multicloud environments, extensive pre-built automation runbooks, and the ability to manage thousands of nodes from a single console
- The unified experience with Amazon Q Developer integration enables natural language queries for faster issue resolution

Service Integration

Systems Manager deeply integrates with core AWS services to provide comprehensive infrastructure management capabilities. Key integrations include Amazon EC2 for instance management, IAM for access control and permissions, Amazon CloudWatch for monitoring and logging, AWS Config for configuration compliance tracking, Amazon EventBridge for event-driven automation, and Amazon S3 for storing patches and logs:

- Additional integrations encompass AWS Security Hub for compliance monitoring, AWS KMS for encryption of sensitive parameters, AWS CloudTrail for API logging, AWS Organizations for multi-account management, AWS Backup for data protection, and AWS Lambda for serverless automation workflows
- These integrations enable seamless operational workflows across the entire AWS ecosystem

Onboarding and Offboarding

Customers begin by ensuring the SSM Agent is installed on their managed nodes (pre-installed on most AWS AMIs) and configuring appropriate IAM roles and policies for Systems Manager access. The Default Host Management Configuration simplifies setup by automatically managing instances without requiring instance profiles:

- Users access the unified Systems Manager console to view and manage nodes across accounts and regions
- Getting started involves identifying unmanaged nodes using the automated diagnosis feature, remediating connectivity issues with one-click runbooks, and then leveraging Systems Manager tools for operational tasks
- The service provides Quick Setup options to configure commonly used features with best practices, streamlining the initial deployment process

Getting Started Guide: https://docs.aws.amazon.com/systems-manager/latest/userguide/what-is-systems-manager.html

Data Removal

Systems Manager data removal involves deleting stored parameters from Parameter Store, removing automation runbooks and documents, clearing inventory data, and terminating active sessions. Organizations can delete OpsItems, maintenance windows, patch baselines, and compliance data through the console or APIs. For managed nodes, customers must remove SSM Agent registration and clean up local registration directories. Multi-account deployments require removing delegated administrator permissions and cleaning up cross-account sharing configurations.

Backup/Restore and Disaster Recovery

Systems Manager provides limited direct backup capabilities but integrates with AWS Backup for comprehensive data protection. Parameter Store supports parameter versioning and cross-region replication for configuration data resilience. Automation runbooks and documents can be exported and recreated across regions. The service stores operational data in durable AWS services like S3 and CloudWatch for persistence. Organizations implement disaster recovery through Infrastructure as Code practices, replicating Systems Manager configurations across regions.

Backup Capabilities

Systems Manager itself doesn't provide traditional backup capabilities but supports data durability through parameter versioning in Parameter Store and integration with AWS Backup for protecting underlying resources. Operational data like inventory information, compliance data, and patch histories are stored in durable AWS services. Customers can export automation documents and runbooks for backup purposes, while session logs and command outputs can be stored in S3 or CloudWatch Logs for retention.

Disaster Recovery

Systems Manager supports disaster recovery through its distributed architecture across AWS regions and integration with AWS disaster recovery services. The service configuration can be replicated across regions using Infrastructure as Code approaches:

- Parameter Store supports cross-region parameter sharing, and automation runbooks can be deployed to multiple regions
- While not directly integrated with AWS Elastic Disaster Recovery, Systems Manager can be used to orchestrate disaster recovery workflows and automate recovery procedures through custom runbooks

Recovery Objectives

Systems Manager helps achieve RPO and RTO objectives through automation capabilities that enable rapid infrastructure provisioning and configuration deployment across regions. Automation runbooks can orchestrate disaster recovery procedures, reducing manual intervention and recovery time. Parameter Store's cross-region capabilities ensure configuration data availability during regional outages. The service's ability to manage nodes across multiple regions supports active-passive and active-active disaster recovery architectures, helping organizations meet stringent recovery targets through automated processes.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/general/latest/gr/ssm.html

FAQ: https://aws.amazon.com/systems-manager/faq/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/systems-manager/

User Guide: https://docs.aws.amazon.com/systems-manager/latest/userguide/

API Reference: https://docs.aws.amazon.com/systems-manager/latest/APIReference/

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/systems-manager/pricing/

Cost Optimization

Systems Manager offers substantial cost optimization opportunities through automation that reduces manual operational overhead and eliminates the need for bastion hosts, VPNs, and dedicated management infrastructure. The service provides free features including Run Command, Session Manager, State Manager, and Inventory management for basic operations:

- Pay-per-use pricing applies only to advanced features like Just-in-time node access, OpsCenter operations, advanced Parameter Store parameters, and Automation steps
- Organizations can optimize costs by leveraging the free tier for standard operations, using automation to reduce manual labor costs, and eliminating dedicated management infrastructure through the unified console approach

Savings Considerations

Primary cost savings come from operational efficiency gains through automation, reduced need for dedicated management infrastructure, and elimination of bastion hosts and VPN solutions. The service's extensive free tier covers most basic operational needs including session management, command execution, and configuration management:

- Advanced cost savings include reduced personnel costs through automation of routine tasks, faster incident resolution through centralized operations, and improved compliance reducing audit costs
- Organizations save on infrastructure by replacing traditional management solutions with Systems Manager's serverless approach, while automation capabilities reduce both operational overhead and human error-related costs significantly

Service Name

AWS Transfer Family

Service Overview

AWS Transfer Family is a fully managed file transfer service that enables secure transfer of files into and out of AWS storage services including Amazon S3 and Amazon EFS. The service supports multiple protocols including SFTP, FTPS, FTP, AS2, and web browser-based transfers. It maintains existing client-side configurations for authentication, access, and firewalls, ensuring seamless migration of file transfer workflows without requiring changes for customers, partners, or applications. The service provides elastic infrastructure with automatic scaling, high availability across multiple Availability Zones, and integration with native AWS services for security, monitoring, and compliance.

Key Use Cases

- Data lakes: Ingesting external data into Amazon S3 for analytics and machine learning
- B2B file exchanges: Secure business-to-business document exchange using AS2 protocol with trading partners
- Content management: Web serving applications and content distribution workflows
- Compliance-driven workflows: Meeting regulatory requirements for secure file transfers with audit trails
- Supply chain management: Automated file processing and distribution across business units
- Internal data transfers: Moving files between on-premises systems and AWS cloud storage
- Subscription-based data distribution: Automated delivery of data to multiple subscribers

Features

- **Multi-protocol Support**: Supports SFTP, FTPS, FTP, AS2, and web browser-based transfers
- **Managed Workflows**: Automates file processing tasks like decryption, tagging, and Lambda function execution
- **Multiple Authentication Methods**: Service Managed, AWS Directory Service, custom identity providers via Lambda/API Gateway
- **Storage Integration**: Native integration with Amazon S3 and Amazon EFS
- **High Availability**: Multi-AZ deployment with automatic failover and elastic scaling

- **SFTP Connectors**: Bidirectional file transfers between AWS and remote SFTP servers
- **Web Applications**: Browser-based file transfer interface for end users
- **Security Features**: Encryption in transit and at rest, PGP encryption, malware protection

Value Proposition

AWS Transfer Family eliminates the need to manage and operate file transfer infrastructure while providing enterprise-grade security and compliance. It offers seamless migration with zero application changes required, maintaining existing client configurations and workflows:

- The service provides automatic scaling to handle dynamic workloads, 24/7 availability with multi-AZ redundancy, and integration with AWS native services for monitoring, security, and compliance
- Customers benefit from reduced deployment time, eliminated infrastructure costs, simplified management through a single pane of glass, and pay-as-you-use pricing with no upfront costs
- The fully managed nature means no server patching, maintenance, or capacity planning is required

Service Integration

AWS Transfer Family deeply integrates with core AWS services for comprehensive functionality. Primary storage integration includes Amazon S3 for object storage and Amazon EFS for file systems:

- Security integrations encompass AWS IAM for access control, AWS KMS for encryption key management, AWS Secrets Manager for credential storage, and AWS Directory Service for Active Directory integration
- Monitoring and logging leverage Amazon CloudWatch for metrics and alarms, and AWS CloudTrail for API auditing
- Automation capabilities integrate with AWS Lambda for custom processing, Amazon SNS for notifications, and AWS Step Functions for workflow orchestration
- Additional integrations include Amazon API Gateway for custom identity providers, Amazon Cognito for user authentication, and AWS B2B Data Interchange for EDI workflows

Onboarding and Offboarding

Getting started with AWS Transfer Family involves creating an Amazon S3 bucket or EFS file system, configuring an IAM role with appropriate permissions, and creating a Transfer Family server through the AWS Console or APIs. Users must select protocols (SFTP, FTPS, FTP, AS2), configure authentication method (Service Managed, Directory Service, or

Custom), set up user access and home directories, and optionally configure managed workflows for file processing automation:

- The service provides tutorials for common scenarios including SFTP server setup, AS2 configuration, and web application deployment
- No infrastructure provisioning or server management is required, enabling rapid deployment within minutes of configuration completion

Getting Started Guide: https://docs.aws.amazon.com/transfer/latest/userguide/tutorials-transfer.html

Data Removal

Data removal involves deleting Transfer Family servers, users, and associated resources through the AWS Console or APIs. Files stored in Amazon S3 or EFS remain until explicitly deleted by the customer. Managed workflows, connectors, and certificates must be removed separately. Complete offboarding requires deleting IAM roles, policies, and any custom identity provider configurations. CloudWatch logs and metrics can be retained or deleted based on customer requirements. All server endpoints are immediately decommissioned upon server deletion.

Backup/Restore and Disaster Recovery

AWS Transfer Family leverages underlying AWS storage services for backup and disaster recovery. Amazon S3 provides built-in durability with cross-region replication and versioning capabilities. Amazon EFS supports automatic backups and point-in-time recovery. Multi-AZ deployment ensures high availability within regions. Customers can implement cross-region disaster recovery by replicating data and recreating Transfer Family servers in alternate regions. Server configurations can be backed up through Infrastructure as Code templates for rapid recovery.

Backup Capabilities

Transfer Family itself does not provide native backup capabilities but relies on underlying storage services. Amazon S3 offers versioning, cross-region replication, and integration with AWS Backup for centralized backup management:

- Amazon EFS provides automatic backups, point-in-time recovery, and AWS Backup integration
- Server configurations and managed workflows can be exported and version-controlled through CloudFormation or Terraform templates for configuration backup and recovery

Disaster Recovery

AWS Transfer Family supports disaster recovery through multi-AZ deployment within regions and cross-region replication capabilities. The service integrates with AWS Application Recovery Controller for automated failover orchestration:

- Customers can implement warm standby or active-active architectures across regions
- Server endpoints can be quickly recreated in alternate regions using Infrastructure as Code
- DNS routing with Route 53 enables automatic traffic redirection during regional failures

Recovery Objectives

AWS Transfer Family helps meet RPO objectives through real-time data replication to Amazon S3 and EFS, with cross-region replication available for near-zero RPO. RTO objectives are supported through multi-AZ availability ensuring minimal downtime during failures, automated failover capabilities, and Infrastructure as Code for rapid server recreation. Managed workflows can be configured for immediate processing, and the elastic infrastructure eliminates capacity constraints that could impact recovery times.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/general/latest/gr/transfer-service.html

FAQ: https://aws.amazon.com/aws-transfer-family/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/transfer/latest/userguide/what-is-aws-transfer-family.html

User Guide: https://docs.aws.amazon.com/transfer/latest/userguide/what-is-aws-transfer-family.html

API Reference: https://docs.aws.amazon.com/transfer/latest/APIReference/api-welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/aws-transfer-family/pricing/

Cost Optimization

AWS Transfer Family offers several cost optimization strategies including protocol-based pricing allowing selection of only needed protocols to minimize hourly charges. Pay-as-you-use model eliminates infrastructure costs and capacity planning expenses:

- SFTP connectors reduce data transfer costs by enabling direct server-to-server transfers without intermediate storage
- Managed workflows eliminate custom infrastructure for file processing
- Multi-protocol support on single servers reduces endpoint proliferation
- Data compression and efficient transfer patterns can reduce data transfer charges
- Integration with S3 Intelligent Tiering and lifecycle policies optimizes long-term storage costs

Savings Considerations

Primary cost savings come from eliminating dedicated file transfer infrastructure, reducing operational overhead through fully managed service model, and avoiding server maintenance costs. Pay-per-use pricing means costs scale with actual usage rather than provisioned capacity:

- Integration with AWS native services eliminates third-party software licensing
- Automated workflows reduce manual processing costs and improve operational efficiency
- Multi-AZ redundancy is included without additional infrastructure investment
- Free tier provides 12 months of limited usage for testing and small workloads
- Data transfer optimization through direct AWS integration reduces bandwidth costs compared to traditional FTP servers

Service Name

AWS Transfer for SFTP

Service Overview

AWS Transfer for SFTP is a fully managed file transfer service that enables secure file transfers using the SFTP (Secure File Transfer Protocol) into and out of AWS storage services. The service operates as part of AWS Transfer Family and supports transferring data to Amazon S3 and Amazon EFS without requiring server infrastructure management. Transfer for SFTP maintains existing client-side configurations for authentication, access, and firewalls, allowing seamless migration of SFTP workloads to AWS without application modifications. The service provides highly available and elastic infrastructure with full redundancy across multiple Availability Zones within an AWS Region.

Key Use Cases

- Business-to-business (B2B) file exchanges: Securely exchanging files with trading partners and external organizations using standardized SFTP protocol
- Data lake ingestion: Uploading files from third parties such as vendors and partners to AWS data lakes for analytics and processing
- Internal data transfers: Moving files within organizations between different departments or systems using familiar SFTP workflows
- Compliance-driven workflows: Meeting regulatory requirements for secure file transfers in industries like financial services, healthcare, and retail
- Subscription-based data distribution: Delivering data to customers and partners on regular schedules through secure SFTP endpoints
- Legacy system migration: Migrating existing SFTP-based workflows to AWS without modifying client applications or user procedures

Features

- **SFTP Protocol Support**: Full support for SFTP version 3 protocol with SSH authentication and encryption
- **Multiple Storage Backends**: Native integration with Amazon S3 and Amazon EFS for data storage
- **Identity Provider Options**: Support for service-managed, AWS Directory Service, custom Lambda, and API Gateway authentication
- **SFTP Connectors**: Automated file transfers between Amazon S3 and remote SFTP servers with bidirectional capabilities
- **VPC Endpoint Support**: Private connectivity through VPC endpoints for secure, internal network access

- **Managed Workflows**: Post-upload file processing including encryption, decryption, tagging, and custom Lambda functions
- **Logical Directories**: Simplified directory structures mapped to Amazon S3 bucket paths
- **Multi-Protocol Support**: Single server can support multiple protocols (SFTP, FTPS, FTP, AS2)

Value Proposition

AWS Transfer for SFTP eliminates the need to manage SFTP server infrastructure while maintaining compatibility with existing client applications and workflows. The service provides automatic scaling, high availability across multiple Availability Zones, and seamless integration with AWS storage services for enhanced analytics and processing capabilities:

- Key advantages include no upfront costs, pay-as-you-go pricing, built-in security features like encryption and access controls, and compliance certifications including FedRAMP and HIPAA eligibility
- The service reduces operational overhead by eliminating server maintenance, patching, and capacity planning while providing enterprise-grade reliability and performance

Service Integration

AWS Transfer for SFTP integrates natively with core AWS services to provide comprehensive file transfer and processing capabilities. Primary storage integrations include Amazon S3 for object storage and Amazon EFS for file system storage:

- Security and access management integrate with AWS IAM for user permissions, AWS Directory Service for Active Directory authentication, and AWS Secrets Manager for credential storage
- Monitoring and logging leverage Amazon CloudWatch for metrics and logs, AWS CloudTrail for API auditing, and Amazon EventBridge for event-driven workflows
- Additional integrations include AWS Lambda for custom processing, AWS KMS for encryption key management, Amazon VPC for network isolation, and AWS Certificate Manager for SSL/TLS certificates

Onboarding and Offboarding

Getting started with AWS Transfer for SFTP involves creating an SFTP-enabled server through the AWS Transfer Family console, selecting identity provider options (service-managed, Directory Service, or custom), and configuring endpoint settings (publicly accessible or VPC-hosted). Users must create an Amazon S3 bucket or Amazon EFS file system for data storage and configure IAM roles with appropriate permissions:

- After server creation, add users with SSH public keys for authentication and configure access permissions
- Testing can be performed using standard SFTP clients like OpenSSH, WinSCP, or Cyberduck
- The service supports both programmatic setup via APIs/CLI and console-based configuration with step-by-step wizards

Getting Started Guide: https://docs.aws.amazon.com/transfer/latest/userguide/getting-started.html

Data Removal

Offboarding from AWS Transfer for SFTP requires stopping the server to prevent new connections, backing up any required data from Amazon S3 or Amazon EFS storage locations, deleting user accounts and associated SSH keys, removing IAM roles and policies created for the service, and finally deleting the Transfer Family server itself. Data stored in Amazon S3 or Amazon EFS must be explicitly deleted as it persists independently of the Transfer Family server. Additional cleanup includes removing VPC endpoints, security groups, and any CloudWatch log groups created for monitoring.

Backup/Restore and Disaster Recovery

AWS Transfer for SFTP leverages the backup and disaster recovery capabilities of its underlying storage services. Data stored in Amazon S3 automatically benefits from S3's 99.999999999% durability and cross-Availability Zone replication. Amazon EFS provides automatic backups and point-in-time recovery options. The Transfer Family service itself is highly available across multiple Availability Zones, eliminating single points of failure. For comprehensive disaster recovery, customers can implement cross-region replication of S3 data and use multiple Transfer Family servers in different regions.

Backup Capabilities

Transfer for SFTP does not have native backup capabilities as it is a file transfer service rather than a data storage service. Backup capabilities depend on the underlying storage services: Amazon S3 provides versioning, cross-region replication, and integration with AWS Backup for automated backup policies. Amazon EFS supports automatic daily backups, point-in-time recovery, and AWS Backup integration for centralized backup management across AWS services.

Disaster Recovery

AWS Transfer for SFTP supports disaster recovery through its multi-Availability Zone architecture and integration with AWS storage services. The service automatically handles failover between Availability Zones without user intervention:

- For broader disaster recovery, customers can deploy Transfer Family servers in multiple AWS regions and implement cross-region data replication using S3 Cross-Region Replication or EFS replication
- The service does not directly integrate with AWS Elastic Disaster Recovery but supports disaster recovery strategies through its storage backends

Recovery Objectives

Transfer for SFTP helps achieve low RTO objectives through its highly available, multi-AZ architecture that automatically recovers from individual Availability Zone failures within minutes. RPO objectives are primarily determined by the underlying storage service backup policies and replication configurations. Amazon S3's cross-region replication can achieve near-zero RPO, while EFS backup policies can provide point-in-time recovery with configurable backup frequencies to balance RPO requirements with cost considerations.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/general/latest/gr/transfer-service.html

FAQ: https://aws.amazon.com/transfer-family/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/transfer/latest/userguide/what-is-aws-transfer-family.html

User Guide: https://docs.aws.amazon.com/transfer/latest/userguide/

API Reference: https://docs.aws.amazon.com/transfer/latest/APIReference/

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/transfer-family/pricing/

Cost Optimization

Transfer for SFTP offers several cost optimization opportunities including paying only for instantiated server hours and data transferred, with no upfront costs or minimum commitments. The service provides a 12-month free tier for new customers:

- Cost optimization strategies include using VPC endpoints to avoid data transfer charges for internal traffic, implementing SFTP connectors to reduce manual transfer operations, rightsizing server configurations based on actual usage patterns, and leveraging S3 storage classes for long-term cost optimization of transferred data

- Monitoring CloudWatch metrics helps identify usage patterns to optimize server uptime and reduce unnecessary costs

Savings Considerations

Primary cost savings come from eliminating the need to manage SFTP server infrastructure, including hardware, operating system licensing, patching, and maintenance costs. The fully managed service reduces operational overhead and staffing requirements while providing enterprise-grade security and compliance capabilities:

- Additional savings include automatic scaling that eliminates over-provisioning, integration with cost-effective AWS storage services, and the ability to implement automated workflows that reduce manual processing costs
- Pay-as-you-go pricing ensures costs align with actual usage, and the service's high availability reduces potential business disruption costs associated with server downtime

Service Name

AWS Transform

Service Overview

AWS Transform is a service that helps accelerate and simplify the transformation of infrastructure, applications, and code with an agentic AI experience. It provides specialized tools to streamline modernization and migration efforts across various workloads including IBM z/OS and Fujitsu GS21 mainframe systems to AWS, VMware workloads to Amazon EC2, .NET applications to Linux-ready cross-platform .NET, and workload assessment for migration readiness. The service offloads labor-intensive, complex transformation tasks across discovery, planning, and execution phases to AI agents with expertise in languages, frameworks, and infrastructure, allowing teams to focus on innovation while efficiently transforming complex, large-scale projects through a unified web experience.

Key Use Cases

- Mainframe modernization: Modernize and migrate IBM z/OS and Fujitsu GS21 mainframe applications to AWS with automated code analysis, business logic extraction, and COBOL to Java conversion
- VMware migration: Migrate VMware workloads to Amazon EC2 with automated application discovery, dependency mapping, migration planning, network conversion, and EC2 instance optimization
- .NET application modernization: Modernize Windows-based .NET applications to Linux-ready cross-platform .NET with automated framework upgrades, package updates, and compatibility conversion
- Full-stack Windows modernization: Coordinated modernization of SQL Server databases, application tiers, and UI frameworks with migration to Amazon Aurora PostgreSQL and deployment to Amazon ECS or EC2 Linux
- Custom transformations: Large-scale modernization of software, code, libraries, and frameworks including API migrations, language version upgrades, framework upgrades, and code refactoring
- Assessment and planning: Assess workloads for migration readiness with automated dependency analysis, complexity assessment, and wave planning for transformation projects

Features

- **Agentic AI Experience**: Uses specialized AI agents with expertise in languages, frameworks, and infrastructure to automate complex transformation tasks

- **Conversational Interface**: Natural language processing interface where users describe current architecture and transformation goals in everyday language
- **Mainframe Modernization**: Automated code analysis, business logic extraction, code decomposition, and COBOL to Java refactoring for IBM z/OS applications
- **VMware Migration**: AI-driven network architecture conversion, migration plan generation, and automated server rehosting to Amazon EC2
- **.NET Modernization**: Automated porting from .NET Framework to cross-platform .NET with Linux compatibility and package modernization
- **Full-stack Windows Modernization**: Coordinated modernization of SQL Server databases, application code, and UI frameworks with automated deployment
- **Custom Transformations**: Large-scale code transformations including API migrations, language upgrades, framework migrations, and organization-specific transformations
- **Wave Planning**: Intelligent dependency analysis and migration wave planning using graph neural networks to optimize transformation sequences
- **Human-in-the-Loop Collaboration**: Unified web experience for cross-functional teams with collaboration features and approval workflows for critical actions
- **Infrastructure as Code Generation**: Automated generation of AWS CloudFormation templates and deployment configurations for target environments

Value Proposition

AWS Transform delivers unprecedented speed and efficiency in enterprise workload modernization through agentic AI automation. The service accelerates transformation projects up to 4x faster than traditional methods while reducing modernization effort and development time:

- For .NET modernization, it can modernize applications up to four times faster with up to 40% savings in licensing costs
- For mainframe modernization, it decomposes monolithic z/OS COBOL applications into cloud-ready components in minutes instead of months
- For VMware migration, it automates network configuration conversion up to 80 times faster than manual approaches
- The service combines advanced AI technologies including foundational models, LLMs, machine learning, graph neural networks, and automated reasoning to handle complex, labor-intensive transformation tasks while preserving critical business logic and maintaining functional equivalence

Service Integration

AWS Transform integrates deeply with core AWS infrastructure and migration services to provide comprehensive transformation capabilities. For compute and deployment, it

works with Amazon EC2, Amazon ECS, and AWS Application Migration Service for server migration and application hosting:

- For database modernization, it integrates with Amazon Aurora PostgreSQL, Amazon RDS, and AWS Database Migration Service for schema conversion and data migration
- For storage and artifacts, it leverages Amazon S3 for secure storage of transformation artifacts and code repositories
- For networking, it uses Amazon VPC, security groups, and AWS PrivateLink for secure connectivity
- For infrastructure management, it generates AWS CloudFormation templates and integrates with AWS Systems Manager for operational management
- The service also works with AWS Identity and Access Management for security, AWS Key Management Service for encryption, Amazon CloudWatch for monitoring, and AWS Application Discovery Service for dependency mapping and assessment

Onboarding and Offboarding

Getting started with AWS Transform involves enabling the service in the AWS Management Console, configuring IAM Identity Center or third-party identity providers, and setting up encryption preferences. Users can choose to enable the command line interface for custom transformations and the web application for the agentic user interface:

- After enabling the service, customers create workspaces which serve as permissions boundaries and resource containers
- The setup process includes configuring connectors to access customer-owned resources in external systems, establishing account connections for AWS or remote services like GitHub, and defining user roles with appropriate permissions (Administrator, Approver, Contributor, or Reader)
- The service provides getting started guides and integrates with AWS Skill Builder for learning modules and AWS Countdown Premium for expert support during the transformation journey

Getting Started Guide: https://docs.aws.amazon.com/transform/latest/userguide/getting-started.html

Data Removal

AWS Transform data removal involves deleting workspaces, connectors, and jobs through the service console or APIs. Users with Administrator permissions can delete workspaces which contain other resources like connectors and jobs. When offboarding, customers should delete transformation artifacts stored in their Amazon S3 buckets, remove connectors that access external resources, and clean up any AWS CloudFormation stacks or infrastructure created during transformations. The service provides APIs for

programmatic deletion of resources. Account connections to external services like GitHub should also be revoked during offboarding to ensure complete data removal.

Backup/Restore and Disaster Recovery

AWS Transform provides data protection through standard AWS infrastructure resilience and security measures. Transformation artifacts are stored in Amazon S3 with built-in durability and versioning capabilities. The service uses AWS Key Management Service for encryption of data, conversion history, and artifacts. Workspaces and job metadata are protected through AWS's standard backup and recovery mechanisms. However, AWS Transform primarily focuses on transformation workflows rather than providing dedicated backup and disaster recovery capabilities. Customers should implement their own backup strategies for transformed code and applications using AWS Backup or other AWS services.

Backup Capabilities

AWS Transform does not provide specific backup capabilities as a core service feature. The service stores transformation artifacts, job history, and workspace data using underlying AWS infrastructure which provides standard durability guarantees:

- Customers are responsible for backing up their source code repositories, transformation results, and any generated infrastructure templates
- The service integrates with standard AWS storage services like Amazon S3 which customers can configure with their own backup policies
- AWS Transform does not directly integrate with AWS Backup service for automated backup management

Disaster Recovery

AWS Transform does not provide built-in disaster recovery capabilities or integration with AWS Elastic Disaster Recovery. The service operates within standard AWS regional infrastructure and inherits the resilience characteristics of underlying AWS services:

- Disaster recovery for AWS Transform workloads would need to be implemented by customers using standard AWS disaster recovery patterns such as multi-region deployments, data replication, and infrastructure as code for rapid environment recreation
- Customers should implement their own disaster recovery strategies for critical transformation projects and associated data

Recovery Objectives

AWS Transform does not directly provide RPO and RTO capabilities as it is a transformation service rather than a data protection service. The service helps customers achieve faster recovery objectives indirectly by modernizing applications to cloud-native architectures that can leverage AWS disaster recovery services. Transformed applications

deployed on Amazon EC2, Amazon ECS, or with Amazon Aurora can then utilize AWS Backup, AWS Elastic Disaster Recovery, and other AWS services to meet specific RPO and RTO requirements based on business needs.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/transform/latest/userguide/transform-limits.html

FAQ: https://aws.amazon.com/transform/faq/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/transform/latest/userguide/what-is-service.html

User Guide: https://docs.aws.amazon.com/transform/latest/userguide/getting-started.html

API Reference: https://docs.aws.amazon.com/transform/latest/userguide/security_iam_permissions.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/transform/pricing/

Cost Optimization

AWS Transform provides significant cost optimization opportunities as its core features including assessment and transformation are offered at no additional cost to AWS customers. The service delivers substantial cost savings through automated modernization - .NET modernization can achieve up to 40% savings in licensing costs by migrating from Windows to Linux:

- Mainframe modernization reduces operational costs by moving from expensive mainframe infrastructure to cost-effective cloud services
- VMware migration eliminates virtualization licensing costs and enables optimization through rightsizing on Amazon EC2
- The service reduces labor costs by automating complex, time-intensive transformation tasks that would otherwise require specialized expertise and extended project timelines, delivering completed projects up to 4x faster than traditional methods

Savings Considerations

Primary cost savings from AWS Transform include elimination of expensive legacy infrastructure costs such as mainframe hardware and software licensing, VMware virtualization licenses, and Windows server licensing. The service reduces project costs by accelerating transformation timelines up to 4x faster, minimizing the need for specialized transformation consultants and extended project teams:

- Operational savings come from modernized applications that can leverage AWS cost optimization features like auto-scaling, spot instances, and reserved capacity
- The automated nature of transformations reduces human error and rework costs
- Long-term savings include reduced maintenance costs for legacy systems, improved developer productivity on modern platforms, and the ability to leverage cloud-native services for enhanced efficiency and reduced operational overhead

Service Name

AWS Transit Gateway

Service Overview

AWS Transit Gateway is a fully managed network transit hub that enables customers to interconnect multiple Virtual Private Clouds (VPCs) and on-premises networks through a central gateway using a hub-and-spoke architecture. It simplifies network management by eliminating the need for complex VPC peering relationships or multiple VPN connections, allowing thousands of VPCs to connect to a single gateway. Transit Gateway acts as a regional virtual router that scales elastically based on network traffic, supporting both IPv4 and IPv6 protocols while providing centralized routing configuration and traffic inspection capabilities across multi-account and multi-region environments.

Key Use Cases

- Centralized connectivity: Connect thousands of VPCs and on-premises networks through a single hub, eliminating complex peering relationships and reducing operational overhead
- Multi-account network architecture: Enable secure communication between VPCs across multiple AWS accounts while maintaining centralized network management and routing policies
- Hybrid cloud connectivity: Integrate on-premises networks with AWS infrastructure using VPN connections, Direct Connect, and SD-WAN solutions through Transit Gateway Connect attachments
- Network segmentation: Implement isolated routing domains and traffic inspection using multiple route tables and centralized security appliances in appliance mode
- Multi-region global networks: Connect networks across AWS regions using Transit Gateway peering attachments while keeping traffic on the AWS backbone

Features

- **Hub-and-spoke architecture**: Central gateway connecting thousands of VPCs and on-premises networks with simplified routing
- **Multiple attachment types**: Support for VPC, VPN, Direct Connect, Transit Gateway Connect, peering, and network function attachments
- **Route table management**: Default and custom route tables with dynamic and static routing, route propagation, and traffic segmentation
- **Multicast support**: Native multicast routing with IGMP snooping for broadcasting applications like financial data feeds
- **Cross-region peering**: Inter-region connectivity keeping traffic on AWS backbone with automatic encryption at physical layer

- **Encryption control**: Enforce encryption-in-transit for all traffic on attached VPCs with configurable encryption policies
- **Network monitoring**: Flow logs, CloudWatch metrics, and Network Manager integration for comprehensive network visibility
- **Resource sharing**: AWS RAM integration for sharing transit gateways across multiple accounts with granular permissions
- **High bandwidth performance**: Up to 100 Gbps per VPC attachment per AZ and 20 Gbps per Connect attachment with ECMP support
- **Flexible cost allocation**: Customizable cost allocation policies for data processing charges across attachments and accounts

Value Proposition

AWS Transit Gateway provides significant advantages over traditional networking approaches by dramatically reducing operational complexity through its managed hub-and-spoke model, eliminating the need for hundreds of VPC peering connections. It offers enterprise-grade scalability supporting thousands of VPC connections per gateway while maintaining high performance with up to 100 Gbps bandwidth per attachment:

- The service provides enhanced security through built-in encryption controls and centralized traffic inspection capabilities
- Cost efficiency is achieved through simplified architecture reducing the number of required connections and flexible cost allocation options
- Unlike self-managed alternatives, Transit Gateway is fully managed by AWS, reducing operational overhead while providing native integration with AWS services like Direct Connect, VPN, and Network Manager for comprehensive network visibility and control

Service Integration

Transit Gateway natively integrates with Amazon VPC for attaching virtual private clouds and managing routing. AWS Direct Connect integration enables high-bandwidth hybrid connectivity through Direct Connect gateways:

- Site-to-Site VPN connections provide secure tunnels to on-premises networks with BGP dynamic routing support
- AWS Resource Access Manager (RAM) enables cross-account sharing of transit gateways
- Transit Gateway Connect attachments integrate with third-party SD-WAN and network appliances using GRE tunnels and BGP
- CloudWatch provides monitoring metrics and Network Manager offers centralized network management and visualization

- Integration with Route 53 Resolver enables DNS resolution across connected networks, while AWS Network Firewall can be deployed for centralized security inspection in appliance mode configurations

Onboarding and Offboarding

Customers begin by creating a Transit Gateway in their desired AWS region through the AWS Management Console, CLI, or API, specifying basic configuration like ASN and enabling optional features like DNS support and multicast. Next, they create VPC attachments by selecting subnets in each VPC they want to connect, with Transit Gateway automatically creating elastic network interfaces:

- Route tables are configured to control traffic flow between attachments, with options for route propagation from VPN and Direct Connect connections
- For hybrid connectivity, customers establish VPN connections or Direct Connect gateway attachments
- Testing involves verifying connectivity between resources in different VPCs using standard network tools like ping or application-specific tests
- AWS provides comprehensive tutorials and best practice guides for common scenarios including centralized routing, isolated networks, and multi-region deployments

Getting Started Guide: https://docs.aws.amazon.com/vpc/latest/tgw/tgw-getting-started.html

Data Removal

Offboarding requires systematic detachment of all resources before deletion. First, delete all route table associations and route propagations, then detach VPCs, VPN connections, Direct Connect gateways, and peering connections. Each attachment must be fully detached and reach ‘deleted’ state before proceeding. Finally, delete the Transit Gateway itself, which automatically removes the default route table and any custom route tables. All routing configurations, flow logs, and associated metadata are permanently removed. For shared transit gateways, the owner must revoke resource shares through AWS RAM before deletion.

Backup/Restore and Disaster Recovery

Transit Gateway configurations are not directly backed up by AWS Backup service as it’s infrastructure-as-code. Disaster recovery relies on Infrastructure as Code (IaC) templates using CloudFormation, CDK, or Terraform to recreate configurations. Route table configurations, attachments, and policies should be documented and version-controlled. For multi-region disaster recovery, Transit Gateway peering enables cross-region connectivity with automatic failover through routing updates. Network configurations can

be replicated across regions using automated deployment pipelines. Flow logs and monitoring data can be archived to S3 for historical analysis and compliance requirements.

Backup Capabilities

Transit Gateway does not support traditional backup operations as it's a managed networking service with infrastructure configurations. The service maintains high availability through AWS's redundant infrastructure across multiple Availability Zones:

- Configuration backup is achieved through Infrastructure as Code practices, exporting route tables and attachment configurations to version-controlled templates
- Transit Gateway is not compatible with AWS Backup service since it doesn't store customer data but rather provides network routing services
- Recovery relies on recreating the configuration using saved IaC templates rather than restoring from backup snapshots

Disaster Recovery

Transit Gateway supports disaster recovery through multi-region peering connections that maintain encrypted connectivity across regions on the AWS backbone. It integrates with AWS Elastic Disaster Recovery by providing network connectivity for replicated resources across regions:

- Cross-region peering enables automatic traffic rerouting during regional failures
- The service supports disaster recovery architectures by connecting recovery VPCs to production networks through peered transit gateways
- Route propagation and dynamic routing enable automatic failover scenarios
- However, Transit Gateway itself is not directly managed by AWS Elastic Disaster Recovery service but serves as the network foundation for DR architectures

Recovery Objectives

Transit Gateway supports aggressive RTO and RPO objectives through its multi-Availability Zone architecture providing sub-second failover within regions. Cross-region peering enables RPO of near-zero for network connectivity with RTO measured in minutes for DNS-based failover. Dynamic routing protocols like BGP provide automatic path selection during failures. Network configurations can be restored rapidly through IaC deployment, typically within 5-15 minutes. For critical applications, multi-region active-active architectures using peered transit gateways can achieve RPO/RTO of seconds, while disaster recovery scenarios with pre-configured cross-region peering can meet most enterprise requirements for sub-hour recovery objectives.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/vpc/latest/tgw/transit-gateway-quotas.html

FAQ: https://aws.amazon.com/transit-gateway/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/vpc/latest/tgw/

User Guide: https://docs.aws.amazon.com/vpc/latest/tgw/what-is-transit-gateway.html

API Reference: https://docs.aws.amazon.com/AWSEC2/latest/APIReference/OperationList-query-ec2.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/transit-gateway/pricing/

Cost Optimization

Transit Gateway offers several unique cost optimization features including Flexible Cost Allocation (FCA) that enables custom allocation of data processing charges to source, destination, or central accounts based on defined policies. This addresses common scenarios where central IT teams bear costs for partner connections or security teams pay for inspection services:

- Cost allocation tags provide granular tracking of charges across multiple teams and projects
- VPN Concentrator attachments reduce costs for multiple low-bandwidth remote sites by consolidating up to 100 connections under a single attachment
- Strategic use of Transit Gateway Connect attachments eliminates data processing charges compared to standard VPC attachments
- Multi-region peering keeps inter-region traffic on AWS backbone, potentially reducing data transfer costs compared to internet routing

Savings Considerations

Primary cost savings come from architectural simplification, replacing hundreds of VPC peering connections with a single Transit Gateway, reducing both hourly attachment costs and operational overhead. Centralized NAT gateways and internet gateways reduce duplicate infrastructure across VPCs:

- Consolidated Direct Connect and VPN connections serve multiple VPCs through single attachments rather than per-VPC connections
- Data processing charges can be optimized through careful route table design to minimize unnecessary traffic traversal
- For multi-account environments, shared Transit Gateway through AWS RAM eliminates duplicate gateways per account
- Long-term savings include reduced networking operational complexity, fewer required networking specialists, and simplified troubleshooting through centralized routing and monitoring
- Strategic placement of Transit Gateway in hub accounts optimizes cross-account data charges through careful cost allocation policy design

Service Name

AWS Trusted Advisor

Service Overview

AWS Trusted Advisor is a proactive guidance service that inspects AWS environments and provides real-time recommendations to save money, improve system availability and performance, and help close security gaps. The service draws upon best practices learned from serving hundreds of thousands of AWS customers. Trusted Advisor analyzes resource configurations, usage patterns, and security posture across multiple categories including cost optimization, performance, security, fault tolerance, service limits, and operational excellence. It delivers actionable recommendations through automated checks that continuously monitor AWS environments, helping customers optimize their cloud infrastructure according to AWS Well-Architected Framework principles.

Key Use Cases

- Use case 1: Cost optimization by identifying underutilized resources, rightsizing recommendations, and reserved instance opportunities to reduce AWS spending
- Use case 2: Security enhancement through automated security gap detection, access permission analysis, and compliance monitoring across AWS resources
- Use case 3: Performance optimization by analyzing resource configurations, identifying bottlenecks, and providing recommendations for improved system performance and availability

Features

- **Cost Optimization Checks**: Over 16 integrated checks from Cost Optimization Hub providing customized recommendations for EC2, RDS, Lambda, and other services with accurate cost savings estimates
- **Security Monitoring**: Automated security gap detection including exposed access keys, security group configurations, and IAM policy analysis with 111+ Security Hub CSPM controls
- **Performance Analysis**: Resource utilization analysis and performance recommendations including throughput optimization and availability zone redundancy checks
- **Service Limits Monitoring**: Real-time tracking of AWS service quotas and limits with alerts when approaching thresholds to prevent service disruptions
- **Organizational View**: Centralized visibility across multiple AWS accounts through integration with AWS Organizations for enterprise-wide governance

- **Trusted Advisor Priority**: Prioritized and context-driven recommendations with collaboration features for Enterprise Support customers including multi-account views

Value Proposition

Customers should choose AWS Trusted Advisor because it provides automated, proactive guidance based on AWS best practices learned from hundreds of thousands of customers. Unlike reactive monitoring tools, Trusted Advisor continuously analyzes environments and delivers actionable recommendations across six critical areas:

- The service offers significant value through cost savings identification, security risk mitigation, and performance optimization without requiring manual configuration
- For Enterprise Support customers, Trusted Advisor Priority provides curated, prioritized recommendations with Technical Account Manager collaboration
- The service integrates seamlessly with existing AWS environments and scales across organizations through AWS Organizations integration, making it essential for maintaining optimal cloud operations

Service Integration

AWS Trusted Advisor deeply integrates with core AWS services to provide comprehensive analysis and automation. Key integrations include AWS Organizations for multi-account management and organizational view, Amazon CloudWatch Events and EventBridge for automated notifications and workflow triggers, AWS Config for configuration monitoring, AWS Security Hub for security posture management with 111+ CSPM controls, and AWS Cost Optimization Hub for enhanced cost recommendations:

- Additional integrations encompass AWS Systems Manager for operational insights, AWS Well-Architected Tool for workload optimization, AWS Resilience Hub for fault tolerance checks, and AWS Backup for backup validation
- The service also connects with Amazon SNS for notifications, AWS IAM for access management, and AWS Support API for programmatic access

Onboarding and Offboarding

Customers get started with AWS Trusted Advisor through their AWS Support plan level. Basic Support provides access to Service Limits checks and selected Security/Fault Tolerance checks with manual refresh capabilities:

- Business Support+, Enterprise Support, and Unified Operations customers access all checks automatically through the AWS Management Console at console.aws.amazon.com/trustedadvisor/home
- For Enterprise customers, Trusted Advisor Priority requires account team enablement and AWS Organizations setup

- Customers can immediately view recommendations, set up CloudWatch monitoring, configure notifications, and integrate with existing workflows
- The service requires no installation or configuration, automatically analyzing environments and providing recommendations based on detected resources and configurations across supported AWS services

Getting Started Guide: https://docs.aws.amazon.com/awssupport/latest/user/get-started-with-aws-trusted-advisor.html

Data Removal

AWS Trusted Advisor does not store customer data beyond recommendation metadata and check results. When customers discontinue AWS Support plans or close AWS accounts, Trusted Advisor automatically stops generating new recommendations and removes access to historical check results. No specific data removal process is required as the service only maintains configuration analysis results rather than actual customer data. Customers can disable individual checks, exclude resources from recommendations, or stop using the service entirely by downgrading support plans without data retention concerns.

Backup/Restore and Disaster Recovery

AWS Trusted Advisor does not provide direct backup and disaster recovery capabilities as it is an advisory service. However, it supports backup strategies through integration with AWS Resilience Hub for fault tolerance checks, AWS Backup validation recommendations, and disaster recovery best practice guidance. The service identifies resources lacking backup configurations, monitors backup policies, and provides recommendations for improving recovery readiness. Trusted Advisor helps customers meet RTO and RPO objectives by recommending proper backup configurations and disaster recovery architectures.

Backup Capabilities

AWS Trusted Advisor does not have inherent backup capabilities as it is a guidance service that analyzes configurations rather than storing data. The service does not integrate directly with AWS Backup for its own operations. However, Trusted Advisor provides recommendations for backup strategies, identifies resources without proper backup configurations, and helps validate backup readiness through integration with AWS Resilience Hub and other AWS services.

Disaster Recovery

AWS Trusted Advisor does not directly support disaster recovery operations but provides critical guidance for disaster recovery planning. The service does not integrate with AWS Elastic Disaster Recovery for its own operations. Instead, it helps customers build resilient architectures by identifying single points of failure, recommending availability zone

redundancy, and providing fault tolerance checks through AWS Resilience Hub integration for applications and infrastructure components.

Recovery Objectives

AWS Trusted Advisor helps customers meet RPO and RTO objectives by identifying architectural weaknesses, recommending redundancy configurations, and providing fault tolerance checks. Through integration with AWS Resilience Hub, it monitors application resilience scores and policy breaches. The service recommends backup strategies, multi-AZ deployments, and disaster recovery best practices. While not directly providing recovery capabilities, Trusted Advisor guides customers toward configurations that support faster recovery times and minimize data loss during incidents.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/general/latest/gr/aws_service_limits.html

FAQ: https://aws.amazon.com/premiumsupport/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/awssupport/latest/user/trusted-advisor.html

User Guide: https://docs.aws.amazon.com/awssupport/latest/user/get-started-with-aws-trusted-advisor.html

API Reference: https://docs.aws.amazon.com/awssupport/latest/user/get-started-with-aws-trusted-advisor-api.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/premiumsupport/plans/

Cost Optimization

AWS Trusted Advisor offers unique cost optimization through intelligent resource analysis and integration with AWS Cost Optimization Hub, providing over 16 specialized checks with accurate cost savings estimates. The service identifies underutilized resources, recommends rightsizing opportunities, analyzes reserved instance and savings plan potential, and provides personalized recommendations based on actual usage patterns:

- Advanced features include organizational cost visibility, automated cost anomaly detection integration, and prioritized cost recommendations for Enterprise customers

- Trusted Advisor's cost optimization extends beyond simple resource identification to provide contextual recommendations that consider business requirements and usage patterns

Savings Considerations

The main cost savings considerations for AWS Trusted Advisor center around its inclusion in AWS Support plans rather than separate service charges. The service itself is included with AWS accounts, with expanded features available through higher support tiers:

- Major savings come from implementing Trusted Advisor recommendations including rightsizing over-provisioned resources, eliminating idle resources, optimizing storage configurations, and purchasing appropriate reserved instances or savings plans
- Enterprise customers benefit from Trusted Advisor Priority's curated recommendations and Technical Account Manager guidance
- The service's organizational view helps identify cost optimization opportunities across multiple accounts, potentially delivering significant savings that far exceed support plan costs

Service Name

AWS Verified Access

Service Overview

AWS Verified Access is a Zero Trust security service that enables secure access to corporate applications and infrastructure resources without requiring a VPN. It evaluates each application request in real-time based on user identity, device posture, and contextual policies before granting access. The service supports both HTTP(S) and non-HTTP(S) protocols including TCP, SSH, and RDP, providing comprehensive logging and visibility for security incidents and audits. Verified Access implements policy-based access control using the Cedar policy language to enforce fine-grained security requirements.

Key Use Cases

- VPN replacement: Eliminate traditional VPN infrastructure while maintaining secure remote access to corporate applications and resources
- Zero Trust network access: Implement Zero Trust security architecture with identity-aware, contextual access controls based on user and device trust
- Secure database access: Provide secure access to databases, SAP systems, and other TCP-based applications with continuous monitoring and policy enforcement
- SaaS application security: Extend secure access to SaaS applications and PrivateLink endpoints for users outside AWS environments
- Remote workforce enablement: Support hybrid and remote workforces with secure, device-independent access to corporate resources

Features

- **Zero Trust Access Control**: Evaluates user identity and device posture before granting access with continuous monitoring
- **Multi-Protocol Support**: Supports HTTP(S) and non-HTTP(S) protocols including TCP, SSH, and RDP
- **Device Posture Assessment**: Integrates with device trust providers like CrowdStrike, Jamf, and JumpCloud for device security validation
- **Policy-Based Access**: Uses Cedar policy language for fine-grained, conditional access control based on user attributes and context
- **Identity Provider Integration**: Supports AWS IAM Identity Center, Active Directory, SAML 2.0, and OpenID Connect providers
- **Comprehensive Logging**: Logs every access attempt with detailed visibility for security incidents and compliance audits

- **Connectivity Client**: Desktop application for secure access to non-HTTP resources with encrypted tunneling

Value Proposition

AWS Verified Access offers superior security over traditional VPNs through Zero Trust principles that continuously verify user identity and device posture. Unlike VPNs that provide broad network access, Verified Access grants application-specific access based on contextual policies, reducing attack surface and preventing lateral movement:

- The service eliminates VPN infrastructure complexity while providing comprehensive logging for compliance and security monitoring
- Integration with existing identity providers and device management solutions simplifies deployment, while support for both HTTP and non-HTTP protocols covers diverse application portfolios
- The pay-per-use pricing model with no upfront commitments provides cost advantages over traditional VPN solutions

Service Integration

AWS Verified Access integrates deeply with AWS IAM Identity Center for centralized user authentication and single sign-on capabilities. It leverages Amazon VPC for network connectivity and security groups for traffic control:

- Integration with AWS Certificate Manager provides TLS/SSL certificate management for secure connections
- The service works with Amazon Route 53 for DNS resolution and AWS PrivateLink for secure access to SaaS endpoints
- CloudTrail integration provides comprehensive API logging, while AWS WAF can be attached for additional web application protection
- For non-HTTP access, it integrates with Amazon RDS, EC2 instances, and Network Load Balancers

Onboarding and Offboarding

Getting started requires enabling AWS IAM Identity Center as a trust provider, creating security groups, and having applications behind load balancers with TLS certificates and public domains. Customers create Verified Access instances, trust providers, groups, and endpoints through the AWS console or CLI:

- DNS configuration maps application domains to Verified Access endpoints using CNAME records
- Access policies written in Cedar language define security requirements
- For non-HTTP applications, users install the Connectivity Client on their devices

- The tutorial provides step-by-step guidance through the Amazon VPC console with IAM policies for resource creation

Getting Started Guide: https://docs.aws.amazon.com/verified-access/latest/ug/getting-started.html

Data Removal

Offboarding requires systematic deletion of Verified Access resources in reverse order: endpoints first, then groups, instances, and trust providers. Resources can be deleted through the AWS console or CLI with confirmation prompts requiring typing 'delete'. Associated DNS configurations and client applications should be removed. All access logs are retained according to CloudTrail configuration. Service-linked roles are automatically deleted when the last Verified Access endpoint is removed.

Backup/Restore and Disaster Recovery

AWS Verified Access is a managed service where AWS handles infrastructure backup and disaster recovery. The service configuration including instances, groups, endpoints, and policies can be recreated using Infrastructure as Code through CloudFormation or Terraform. Access logs are stored in CloudTrail for historical audit purposes. For business continuity, organizations should document and version their Cedar policies and trust provider configurations for rapid redeployment across regions.

Backup Capabilities

AWS Verified Access does not have traditional backup capabilities as it is a managed access control service. The service configuration and policies should be backed up through Infrastructure as Code practices. AWS Verified Access does not integrate directly with AWS Backup service since it manages access policies and configurations rather than data that requires backup and restore functionality.

Disaster Recovery

AWS Verified Access supports multi-region deployment for disaster recovery scenarios. The service does not integrate with AWS Elastic Disaster Recovery as it is an access control service rather than a data replication service. Organizations can deploy Verified Access instances across multiple AWS regions and use Route 53 health checks for automatic failover to healthy endpoints during regional outages.

Recovery Objectives

AWS Verified Access supports low RTO objectives through its managed service architecture and multi-region deployment capabilities. Since it is a stateless access control service, recovery primarily involves recreating service configurations rather than data recovery. RPO is not applicable as the service does not store customer data.

Organizations can achieve near-zero RTO by maintaining Infrastructure as Code templates for rapid deployment across regions during disaster scenarios.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/verified-access/latest/ug/verified-access-quotas.html

FAQ: https://aws.amazon.com/verified-access/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/verified-access/latest/ug/what-is-verified-access.html

User Guide: https://docs.aws.amazon.com/verified-access/latest/ug/getting-started.html

API Reference: https://docs.aws.amazon.com/AWSEC2/latest/APIReference/operation-list-verified-access.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/verified-access/pricing/

Cost Optimization

AWS Verified Access offers several cost optimization strategies including pay-per-use pricing with no upfront commitments or minimum fees. For HTTP(S) applications, costs are based on application hours and data processed per GB:

- Non-HTTP applications are charged per endpoint hour with up to 100 free connections per hour
- Organizations can optimize costs by consolidating similar applications into groups to share policies, rightsizing the number of endpoints, and leveraging the free connection allowances
- The elimination of VPN infrastructure reduces overall networking costs while the managed service model eliminates operational overhead expenses

Savings Considerations

Major cost savings include elimination of traditional VPN infrastructure, hardware, and maintenance costs. The managed service model reduces operational expenses for security infrastructure management:

- Pay-per-use pricing ensures customers only pay for actual usage rather than provisioned capacity
- Consolidation of access policies through groups reduces administrative overhead
- The service eliminates costs associated with separate access and connectivity solutions for non-HTTP resources
- Integration with existing AWS services leverages current investments while the serverless architecture provides automatic scaling without capacity planning costs
- Improved security posture reduces potential breach remediation costs through Zero Trust access controls

Service Name

AWS VPN

Service Overview

AWS VPN is comprised of two services: AWS Site-to-Site VPN and AWS Client VPN. AWS Site-to-Site VPN is a fully-managed service that creates secure IPsec VPN connections between on-premises networks or branch offices and AWS resources using IP Security (IPSec) tunnels. AWS Client VPN is a fully-managed remote access VPN solution that enables secure access to resources within both AWS and on-premises networks for remote workforces. Both services provide encrypted connectivity, high availability, and automatic scaling to support hybrid cloud architectures and remote work scenarios.

Key Use Cases

- Hybrid connectivity: Securely extend on-premises networks to AWS cloud resources using Site-to-Site VPN with IPsec tunnels for application migration and multi-site communication
- Remote workforce access: Enable secure remote access for employees to both AWS and on-premises resources using Client VPN with OpenVPN protocol and multi-device support
- Multi-site networking: Connect distributed branch offices and remote locations through VPN CloudHub for hub-and-spoke topology or VPN Concentrator for cost-effective multi-site connectivity

Features

- **Site-to-Site VPN**: IPsec VPN connections with dual tunnels for high availability, supporting up to 5 Gbps per tunnel with accelerated options using AWS Global Accelerator
- **Client VPN**: OpenVPN-based remote access with mutual authentication, SAML/Active Directory integration, and support for multiple operating systems including Windows, macOS, Linux, iOS, and Android
- **VPN Concentrator**: Cost-effective multi-site connectivity allowing up to 100 remote sites per concentrator with shared bandwidth up to 100 Mbps per tunnel
- **Advanced Security**: Certificate-based authentication, Network Address Translation (NAT) traversal, private IP VPN support, and integration with AWS security services
- **Monitoring and Management**: CloudWatch metrics, connection logging, AWS Console management, and integration with AWS monitoring tools for visibility and troubleshooting

Value Proposition

AWS VPN provides fully-managed, highly available VPN services that eliminate the need for customers to deploy and maintain their own VPN infrastructure. Key advantages include automatic scaling and failover, built-in redundancy with dual tunnels, seamless integration with AWS services like Transit Gateway and Direct Connect, and pay-as-you-go pricing with no upfront costs:

- The service offers enterprise-grade security with IPsec encryption, supports high throughput up to 5 Gbps per tunnel, and provides comprehensive monitoring through CloudWatch
- AWS VPN integrates natively with AWS security and networking services, reducing operational complexity while maintaining industry-standard security protocols

Service Integration

AWS VPN integrates deeply with core AWS networking services including Amazon VPC for secure network isolation, AWS Transit Gateway for centralized connectivity hub supporting ECMP routing, and AWS Direct Connect for hybrid encrypted connections. Security integrations include AWS Certificate Manager (ACM) for SSL/TLS certificate management, AWS IAM for authentication and authorization, Amazon CloudWatch for monitoring and metrics, and AWS CloudTrail for API logging. Additional integrations include AWS Global Accelerator for Site-to-Site VPN acceleration, AWS Directory Service for Active Directory authentication with Client VPN, and AWS Network Firewall for traffic inspection at the network perimeter.

Onboarding and Offboarding

Customers can start with AWS VPN using the AWS Management Console, AWS CLI, or APIs. For Site-to-Site VPN, the process involves creating a customer gateway representing the on-premises device, creating a virtual private gateway or Transit Gateway attachment, configuring routing, and downloading configuration files for the customer gateway device:

- For Client VPN, setup requires generating server and client certificates using OpenVPN easy-rsa utility, importing certificates to AWS Certificate Manager, creating a Client VPN endpoint, associating target networks, and configuring authorization rules
- AWS provides comprehensive documentation, sample configurations for supported devices, and getting started guides for both services

Getting Started Guide:
https://docs.aws.amazon.com/vpn/latest/s2svpn/SetUpVPNConnections.html

Data Removal

To offboard from AWS VPN, customers can delete VPN connections, Client VPN endpoints, and associated resources through the AWS Console, CLI, or API. For Site-to-Site VPN, this involves deleting VPN connections, customer gateways, and virtual private gateways. For Client VPN, customers delete endpoints, revoke client certificates, and remove associated target networks. All connection logs stored in CloudWatch Logs and configuration data are automatically removed when resources are deleted, ensuring complete data cleanup during service termination.

Backup/Restore and Disaster Recovery

AWS VPN provides inherent redundancy with dual IPsec tunnels per connection for automatic failover, but does not offer traditional backup/restore capabilities as it's a connectivity service. Configuration data is stored in the AWS control plane with built-in redundancy. For disaster recovery, customers can deploy VPN connections across multiple AWS regions and use multiple tunnel configurations with Transit Gateway ECMP routing for high availability and fault tolerance across geographic regions.

Backup Capabilities

AWS VPN does not integrate with AWS Backup as it is a network connectivity service rather than a data storage service. The service configuration is automatically maintained by AWS with built-in redundancy:

- However, customers should backup their customer gateway device configurations and VPN connection settings outside of AWS
- Client VPN configuration files and certificates should be securely stored and backed up by customers to ensure service continuity

Disaster Recovery

AWS VPN does not integrate with AWS Elastic Disaster Recovery as it provides connectivity rather than compute instance recovery. However, VPN connections support disaster recovery architectures through dual tunnels, multi-region deployments, and integration with Transit Gateway for automatic failover. Customers can establish VPN connections in multiple AWS regions and use routing policies to achieve cross-region disaster recovery for their hybrid network connectivity requirements.

Recovery Objectives

AWS VPN supports low RTO through automatic tunnel failover (typically under 30 seconds), multiple tunnel configurations with ECMP routing, and multi-region VPN deployments. For RPO, the service provides continuous connectivity without data loss during failover events. Accelerated Site-to-Site VPN reduces convergence time by routing through AWS Global

Accelerator edge locations. Customers can achieve sub-minute RTO by implementing proper routing configurations and monitoring tunnel health through CloudWatch metrics.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/vpn/latest/s2svpn/vpn-limits.html

FAQ: https://aws.amazon.com/vpn/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/vpn/

User Guide: https://docs.aws.amazon.com/vpn/latest/s2svpn/VPC_VPN.html

API Reference: https://docs.aws.amazon.com/AWSEC2/latest/APIReference/OperationList-query-vpc.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/vpn/pricing/

Cost Optimization

AWS VPN offers several cost optimization strategies including VPN Concentrator for multi-site deployments that reduces per-site costs by sharing bandwidth across up to 100 sites. Customers can optimize costs by rightsizing tunnel bandwidth (choosing standard 1.25 Gbps vs:

- large 5 Gbps based on actual needs), using unaccelerated VPN for non-critical workloads, and implementing proper routing to minimize data transfer costs
- Terminating VPN on Transit Gateway instead of Virtual Private Gateway enables ECMP routing for better bandwidth utilization
- Customers can also leverage AWS Direct Connect with VPN backup for cost-effective hybrid connectivity while maintaining redundancy

Savings Considerations

Primary cost savings come from eliminating on-premises VPN infrastructure costs including hardware, maintenance, and operational overhead. Pay-as-you-go pricing model eliminates upfront capital expenditure with billing based on connection-hours and data transfer:

- VPN Concentrator reduces costs for multi-site deployments by up to 50% compared to individual connections
- Using standard tunnels instead of accelerated VPN saves on Global Accelerator charges for non-critical workloads
- Customers can achieve additional savings by optimizing data transfer patterns, implementing efficient routing policies, and using Regional Data Transfer pricing
- The fully-managed nature eliminates staffing costs for VPN infrastructure management and maintenance

Service Name

AWS WAF

Service Overview

AWS WAF is a web application firewall service that helps protect web applications from attacks by monitoring HTTP(S) requests to protected resources and allowing, blocking, or counting requests based on defined conditions. The service provides protection against common web exploits like SQL injection, cross-site scripting, and DDoS attacks at the application layer. AWS WAF can be deployed on Amazon CloudFront, Application Load Balancer, API Gateway, AWS AppSync, Amazon Cognito user pools, AWS App Runner, AWS Verified Access, and AWS Amplify applications. The service uses web access control lists (ACLs) containing rules that evaluate requests based on criteria like IP addresses, HTTP headers, HTTP body content, URI strings, query strings, and geographic locations.

Key Use Cases

- Web Application Protection: Shield applications from common attacks like SQL injection, cross-site scripting (XSS), file inclusion, and cross-site request forgery
- Bot Control and Mitigation: Block malicious bots while allowing legitimate ones like search engines and monitoring tools using machine learning and advanced detection
- DDoS Protection: Provide application-layer DDoS protection with rate-based rules and automatic baseline detection to mitigate volumetric attacks
- API Security: Protect REST APIs and GraphQL APIs from abuse, rate limiting, and malicious requests across multiple AWS services
- Fraud Prevention: Prevent account takeover attempts and fraudulent account creation using behavioral analysis and credential intelligence
- Geographic Access Control: Block or allow traffic based on geographic location to comply with regional regulations or business requirements

Features

- **Web Traffic Filtering**: Create rules based on IP addresses, HTTP headers, HTTP body, URI strings, and query parameters
- **AWS Managed Rules**: Pre-configured rule groups maintained by AWS for common threats like OWASP Top 10
- **Bot Control**: Advanced bot detection and mitigation using machine learning to distinguish legitimate from malicious bots
- **Rate-based Rules**: Limit the number of requests from specific IP addresses within defined time windows

- **Fraud Control**: Account takeover prevention and account creation fraud detection with behavioral analysis
- **Real-time Visibility**: Real-time metrics, sampled web requests, and comprehensive logging with CloudWatch integration
- **Custom Response**: Configure custom response codes, headers, and error pages when blocking requests
- **Geographic Blocking**: Allow or block requests based on country of origin using GeoIP database

Value Proposition

AWS WAF offers comprehensive application-layer security with native integration across AWS services, eliminating the need for separate security appliances or complex configurations. The service provides expert-level protection through continuously updated managed rules, reducing the security expertise burden on customers:

- With simplified configuration workflows and pre-configured protection packs, setup time is reduced by up to 80%
- AWS WAF scales automatically without capacity planning, offers pay-as-you-use pricing with no upfront costs, and provides unified management through AWS Firewall Manager
- The service integrates seamlessly with AWS Shield for DDoS protection and provides real-time visibility and monitoring capabilities through CloudWatch

Service Integration

AWS WAF integrates natively with multiple AWS services as the first line of defense. Core integrations include Amazon CloudFront for global edge protection, Application Load Balancer for regional application protection, Amazon API Gateway for REST API security, and AWS AppSync for GraphQL API protection:

- Additional integrations cover Amazon Cognito user pools for authentication security, AWS App Runner for container applications, AWS Verified Access for zero-trust access, and AWS Amplify for hosted web applications
- The service works closely with AWS Shield Advanced for enhanced DDoS protection, AWS Firewall Manager for centralized management, Amazon CloudWatch for monitoring and alerting, and supports logging to Amazon S3, CloudWatch Logs, and Amazon Data Firehose

Onboarding and Offboarding

Customers can get started with AWS WAF through the simplified new console experience by creating a protection pack (web ACL) tailored to their application type. The setup process involves selecting AWS resources to protect, choosing initial protection rules from managed rule groups or custom rules, and configuring default actions and logging:

- The new console provides guided workflows with pre-configured protection packs for different application types like e-commerce and transaction processing
- Customers can also use Infrastructure as Code tools like AWS CloudFormation or AWS CDK for programmatic deployment
- The service offers one-click integration with supported AWS services, and customers can start with count mode to test rules before enabling blocking actions

Getting Started Guide: https://docs.aws.amazon.com/waf/latest/developerguide/setup-iap-console.html

Data Removal

To offboard from AWS WAF, customers must first disassociate web ACLs from all protected resources, then delete the web ACLs, rules, rule groups, IP sets, and regex pattern sets. Configuration data is removed immediately upon deletion. Log data stored in CloudWatch Logs, S3, or Data Firehose must be deleted separately according to the respective service's data retention policies. Customers should ensure all AWS resources are disassociated before deletion to prevent service disruptions.

Backup/Restore and Disaster Recovery

AWS WAF configurations are automatically replicated within the region for high availability but do not provide built-in cross-region backup capabilities. Customers must implement their own backup strategies using Infrastructure as Code tools like CloudFormation or by exporting configurations via APIs. Rule configurations, IP sets, and web ACL settings can be recreated programmatically. Log data stored in associated services follows those services' backup and disaster recovery capabilities.

Backup Capabilities

AWS WAF does not provide native backup capabilities for configuration data. Customers must use Infrastructure as Code practices with AWS CloudFormation, AWS CDK, or direct API calls to maintain configuration backups:

- The service automatically maintains rule evaluation state but does not backup historical configurations
- Rule sets and web ACL configurations should be version controlled externally for backup purposes

Disaster Recovery

AWS WAF supports disaster recovery through multi-region deployment strategies where customers deploy identical configurations across regions. The service integrates with Application Recovery Controller for coordinated failover:

- Configuration replication must be managed manually or through automation

- AWS WAF does not integrate directly with AWS Elastic Disaster Recovery but can be part of broader disaster recovery architectures

Recovery Objectives

AWS WAF helps achieve aggressive RTO and RPO objectives through its real-time processing capabilities and regional availability. Rule changes propagate within minutes, and the service maintains sub-second latency for request processing. For disaster recovery scenarios, customers can achieve RTOs of minutes to hours depending on automation sophistication. RPO is near-zero for active configurations since rules are evaluated in real-time without data persistence requirements.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/waf/latest/developerguide/limits.html

FAQ: https://aws.amazon.com/waf/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/waf/latest/developerguide/waf-chapter.html

User Guide: https://docs.aws.amazon.com/waf/latest/developerguide/waf-chapter.html

API Reference: https://docs.aws.amazon.com/waf/latest/APIReference/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/waf/pricing/

Cost Optimization

AWS WAF offers several unique cost optimization strategies including scope-down statements to limit advanced rule evaluation to specific pages, reducing Bot Control costs. Using AWS Shield Advanced eliminates additional WAF charges for high-volume workloads:

- Customers can optimize by applying lower-cost rules before expensive managed rules, separating static content to different distributions, and using protection packs to avoid individual rule charges
- Rate-based rules provide cost-effective DDoS protection compared to always-on blocking rules
- The service offers request-based pricing with no upfront costs, allowing precise cost control based on actual usage

Savings Considerations

Primary cost savings come from preventing application downtime and data breaches through proactive threat blocking. Automated managed rules reduce operational overhead compared to manual security management:

- The pay-per-use model eliminates capital expenses for security appliances
- Integration with AWS Shield Advanced provides additional WAF features at no extra cost for qualifying workloads
- Bot Control helps reduce infrastructure costs by blocking malicious traffic before it reaches applications
- Customers save on operational costs through centralized management with AWS Firewall Manager and reduced need for specialized security expertise through pre-configured protection packs

Service Name

AWS Wavelength

Service Overview

AWS Wavelength is an edge computing infrastructure service that deploys standard AWS compute and storage services to the edge of communications service providers' (CSP) 5G networks. It enables developers to build applications requiring ultra-low latency or edge resiliency by extending Virtual Private Cloud (VPC) to one or more Wavelength Zones. These zones are logical extensions of AWS Regions, managed by the control plane in the Region, allowing applications to serve mobile devices and end users with minimal network latency while maintaining seamless connectivity to full AWS Regional services.

Key Use Cases

- Real-time gaming: Game streaming with ultra-low latency to preserve user experience on devices with limited processing power
- Augmented and Virtual Reality: Reducing Motion to Photon (MTP) latencies for realistic AR/VR experiences in locations without local system servers
- Connected vehicles: Enabling Cellular Vehicle-to-Everything (C-V2X) functionality for intelligent driving, real-time HD maps, and autonomous driving capabilities
- Smart factories: Industrial automation using machine learning inference at the edge for quality control on assembly lines without expensive on-premise GPU servers
- Media and entertainment: Live streaming high-resolution video and audio with interactive experiences and real-time video analytics
- Healthcare: Medical training through mobile games, simulations, and advanced procedures using remote rendering engines
- Online betting and regulated industries: Edge resiliency for data residency requirements within state or country borders

Features

- **Data Sovereignty by Design**: Built according to AWS sovereign-by-design principles providing verifiable control over data location using AWS Nitro System architecture with hardware-based security and isolation
- **Compute Services**: Selection of EC2 instance families (T3, R5, M5, C5, G4) including g4dn.2xlarge with NVIDIA T4 GPUs for ML inference and small-scale training, plus EC2 Auto Scaling
- **Storage Services**: Amazon EBS gp2 volumes for persistent block storage with snapshot and restore capabilities, plus volume size increases without performance impact

- **Container Services**: Support for Amazon EKS and ECS clusters to run Kubernetes and containerized workloads in Wavelength Zones
- **Load Balancing**: Application Load Balancer available in select Wavelength Zones for distributing traffic across multiple targets
- **Carrier Gateway**: Enables connectivity from Wavelength subnet to CSP network, internet, and AWS Region while filtering malicious threats
- **Regional Service Access**: Seamless connection to full range of AWS Regional services like S3 and DynamoDB through same APIs over high-bandwidth network backbone
- **AWS Management Tools**: Support for familiar AWS tools including CloudFormation, CloudWatch, CloudTrail, Systems Manager, and Cost Explorer

Value Proposition

AWS Wavelength uniquely positions compute resources at the edge of 5G networks within telecommunications providers' facilities, delivering ultra-low latency that cannot be achieved with traditional cloud deployments. Unlike other edge solutions, Wavelength offers the full AWS ecosystem with consistent APIs, tools, and services while meeting data residency requirements:

- The service eliminates the need for expensive on-premises infrastructure by leveraging carrier partnerships, providing automatic scaling, built-in security, and seamless integration with Regional AWS services
- This makes it ideal for latency-critical applications like real-time gaming, AR/VR, autonomous vehicles, and industrial automation where milliseconds matter

Service Integration

AWS Wavelength integrates seamlessly with core AWS services through high-bandwidth connections to parent AWS Regions. Primary integrations include Amazon EC2 instances and Auto Scaling, Amazon EBS volumes, Amazon VPC subnets and security groups, Amazon EKS and ECS for containers, and Application Load Balancer in select zones:

- Regional services like Amazon S3, DynamoDB, and RDS are accessible through the same APIs and SDKs
- Management services include Amazon CloudWatch for monitoring, AWS CloudTrail for auditing, AWS CloudFormation for infrastructure as code, and AWS Systems Manager for instance management
- The carrier gateway component provides secure connectivity to telecommunications networks while maintaining access to the full AWS ecosystem

Onboarding and Offboarding

Customers begin by opting into Wavelength Zones through the EC2 Dashboard, ensuring they can create necessary resources in the desired Region and reviewing quotas. The

setup process involves creating a VPC, configuring a carrier gateway, and establishing subnets in both Availability Zones and Wavelength Zones:

- Instances can be launched in Wavelength zones with Carrier IP addresses either auto-assigned or manually allocated
- Network configuration includes security group rules and route table customization
- Testing connectivity between Regional and Wavelength instances, as well as carrier network connectivity, validates the deployment
- The process uses familiar AWS tools including Management Console, AWS CLI, and SDKs for consistent operational experience

Getting Started Guide:
https://docs.aws.amazon.com/wavelength/latest/developerguide/get-started-wavelength.html

Data Removal

Data removal from AWS Wavelength follows standard AWS practices for associated services. Customers terminate EC2 instances, delete EBS volumes and snapshots, remove VPC components including subnets and carrier gateways, and delete any container clusters. Since Wavelength integrates with Regional services, data stored in services like S3 or DynamoDB must be removed from those services separately. All resources are managed through standard AWS APIs and console interfaces, ensuring consistent deletion processes across the platform.

Backup/Restore and Disaster Recovery

AWS Wavelength relies on standard AWS backup mechanisms for each integrated service. EBS volumes support snapshots stored in the parent AWS Region, providing point-in-time recovery capabilities. AMIs are also stored regionally for instance recovery. However, Wavelength itself does not provide specific disaster recovery features - customers must architect resilient solutions using multiple Wavelength Zones, Regional failover patterns, or integration with services like AWS Backup for centralized backup management across their infrastructure.

Backup Capabilities

AWS Wavelength supports backup through its constituent services rather than as a dedicated backup service. EBS volumes in Wavelength Zones can be backed up using EBS snapshots stored in the parent Region:

- The service works with AWS Backup for centralized backup policy management where applicable
- AMIs created from Wavelength instances are stored regionally, enabling instance recovery and replication across zones and regions as needed

Disaster Recovery

AWS Wavelength does not directly integrate with AWS Elastic Disaster Recovery as a standalone service. Instead, disaster recovery must be architected using multiple Wavelength Zones within VPCs, combining with Regional resources, or implementing hub-and-spoke models with the parent Region for failover. Customers can design highly available architectures by distributing applications across multiple Wavelength Zones and implementing DNS-based failover using Route 53 health checks.

Recovery Objectives

AWS Wavelength helps customers meet RPO and RTO objectives through its integration with AWS Regional services and backup mechanisms. EBS snapshots provide point-in-time recovery for data protection, while AMI storage in Regions enables rapid instance deployment. Customers can design architectures with multiple Wavelength Zones for redundancy, implement automated scaling for availability, and use DNS failover for rapid traffic redirection to meet specific recovery time requirements.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/wavelength/latest/developerguide/wavelength-quotas.html

FAQ: https://aws.amazon.com/wavelength/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/wavelength/latest/developerguide/what-is-wavelength.html

User Guide: https://docs.aws.amazon.com/wavelength/latest/developerguide/get-started-wavelength.html

API Reference: https://docs.aws.amazon.com/AWSEC2/latest/APIReference/

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/wavelength/pricing/

Cost Optimization

AWS Wavelength offers unique cost optimization through on-demand pricing models with EC2 Instance Savings Plans applicability, eliminating upfront infrastructure investments

typically required for edge deployments. Cost efficiency comes from leveraging shared telecommunications infrastructure rather than deploying dedicated edge hardware:

- Customers can optimize costs by right-sizing instance types based on workload requirements, using the limited but sufficient instance family options available, and architecting applications to minimize data transfer between Wavelength Zones and Regions
- The hub-and-spoke model with Regional services allows cost-effective resource sharing while maintaining edge performance benefits

Savings Considerations

Primary cost savings stem from avoiding expensive on-premises edge infrastructure deployment and maintenance costs, particularly for GPU-intensive workloads using g4dn instances instead of dedicated hardware. Shared carrier infrastructure reduces operational overhead while providing enterprise-grade connectivity:

- Instance Savings Plans can reduce compute costs, and strategic architectural decisions like processing data locally in Wavelength Zones before sending results to Regional services minimize data transfer costs
- The elimination of physical hardware procurement, maintenance contracts, and dedicated facility requirements for edge computing represents significant capital and operational expenditure reductions for latency-sensitive applications

Service Name

AWS Well-Architected Tool

Service Overview

AWS Well-Architected Tool (AWS WA Tool) is a free service that helps users document and measure their workloads against AWS architectural best practices. It provides recommendations for improving workloads' reliability, security, efficiency, cost-effectiveness, and sustainability. Based on the AWS Well-Architected Framework developed by AWS Solutions Architects, the tool evaluates architectures against five core pillars: Operational Excellence, Security, Reliability, Performance Efficiency, and Cost Optimization. Users can create custom lenses to measure workloads against their specific best practices and conduct reviews across their portfolio of applications.

Key Use Cases

- Architecture Review: Conduct comprehensive reviews of workloads against AWS best practices across the five pillars to identify risks and improvement opportunities
- Product Launch Governance: Document architectures and provide governance for new product launches to ensure alignment with organizational standards
- Technology Portfolio Risk Management: Manage risks across technology portfolios by systematically evaluating multiple workloads and tracking improvements
- Compliance and Audit: Support regulatory and compliance requirements with built-in compliance checks and standardized assessment frameworks
- Continuous Improvement: Track workload improvements over time using milestones and measurement capabilities for ongoing architecture optimization

Features

- **Workload Reviews**: Conduct detailed reviews against AWS Well-Architected Framework pillars with customizable lenses and question sets
- **Custom Lenses**: Create industry-specific or custom evaluation criteria to supplement standard AWS best practices
- **Trusted Advisor Integration**: Access relevant Trusted Advisor checks integrated within workload reviews for comprehensive recommendations
- **Milestone Tracking**: Save workload states at specific points in time to track improvements and changes over time
- **Improvement Plans**: Generate prioritized improvement plans with step-by-step guidance for addressing identified risks
- **Collaboration Features**: Share workloads and reviews with team members, AWS accounts, and AWS Organizations for collaborative assessment

Value Proposition

AWS Well-Architected Tool provides unique value as a free, comprehensive architecture assessment platform directly integrated with AWS services and backed by Solutions Architect expertise. It offers standardized evaluation against proven best practices, enabling consistent architecture reviews across organizations:

- The tool delivers actionable improvement plans with prioritized recommendations, reducing the complexity of optimization initiatives
- Integration with Trusted Advisor and Service Catalog AppRegistry provides contextual insights specific to actual AWS resources
- Custom lens capabilities allow organizations to incorporate industry-specific or proprietary standards alongside AWS best practices, making it adaptable to diverse business requirements while maintaining standardized assessment methodologies

Service Integration

AWS Well-Architected Tool integrates natively with several core AWS services to enhance assessment capabilities. AWS Trusted Advisor integration provides real-time infrastructure recommendations and checks within workload reviews for Business and Enterprise Support customers:

- AWS Service Catalog AppRegistry integration enables workload definition based on existing application resources and metadata
- AWS Organizations integration simplifies resource sharing and management across organizational accounts through service-linked roles
- AWS Identity and Access Management (IAM) provides granular access control with managed policies for full access, read-only access, and organizational permissions
- The tool also connects with AWS Cost Explorer through its cost optimization recommendations and works with AWS Config for additional configuration insights

Onboarding and Offboarding

Getting started with AWS Well-Architected Tool involves setting up appropriate IAM permissions for users, groups, or roles, then activating support for desired AWS service integrations like Trusted Advisor or Service Catalog AppRegistry. Users define workloads by specifying name, description, owner, environment, regions, and accounts:

- The review process involves answering questions across the six framework pillars, with the tool providing improvement plans based on responses
- Organizations can save milestones to track progress and share workloads for collaborative reviews
- The service is accessible through the AWS Management Console in all AWS regions without additional setup requirements, making onboarding straightforward for existing AWS customers

Getting Started Guide:
https://docs.aws.amazon.com/wellarchitected/latest/userguide/getting-started.html

Data Removal

AWS Well-Architected Tool stores workload definitions, review responses, custom lenses, and milestone data within customer AWS accounts. For offboarding, customers can delete individual workloads, custom lenses, and review templates through the console or API. All associated data including answers, milestones, and improvement plans are permanently removed when workloads are deleted. Organizations can also revoke sharing permissions and disable service integrations with Trusted Advisor or Organizations. The service follows standard AWS data retention and deletion practices for customer account closure scenarios.

Backup/Restore and Disaster Recovery

AWS Well-Architected Tool does not provide traditional backup and disaster recovery capabilities as it is a assessment and documentation service rather than a data storage service. The tool itself operates as a managed AWS service with built-in resilience across AWS regions. Workload definitions, reviews, and custom lenses are stored within the regional service infrastructure. Customers can export reports and documentation for external backup purposes, but the service focuses on architecture assessment rather than data protection. Recovery capabilities are inherent to the managed service design rather than customer-configurable backup solutions.

Backup Capabilities

AWS Well-Architected Tool does not have traditional backup capabilities and does not integrate with AWS Backup service since it operates as a managed assessment platform rather than a data storage solution. The service stores configuration data, workload definitions, and assessment results within AWS's managed infrastructure:

- Customers can export reports and assessments for external record-keeping, but the service relies on AWS's underlying infrastructure resilience rather than customer-managed backup solutions
- Data persistence is handled through the managed service architecture across AWS regions

Disaster Recovery

AWS Well-Architected Tool does not support traditional disaster recovery and does not integrate with AWS Elastic Disaster Recovery as it is a managed assessment service rather than infrastructure requiring customer-managed DR capabilities. The service operates with built-in resilience as part of AWS's managed service architecture:

- Regional availability ensures service continuity, with workloads and assessments accessible from supported AWS regions

- The tool actually helps customers design disaster recovery strategies for their workloads through reliability pillar assessments rather than requiring DR configuration for itself

Recovery Objectives

AWS Well-Architected Tool helps customers meet RPO and RTO objectives by providing structured guidance through the Reliability pillar questions and best practices rather than offering direct recovery capabilities. The tool evaluates backup strategies, disaster recovery plans, fault tolerance designs, and business continuity preparations within workload assessments. Through improvement plans, it recommends specific AWS services and architectural patterns that support customer-defined recovery objectives. The assessment framework includes questions about backup automation, cross-region replication, and recovery testing to ensure architectures align with target RPO and RTO requirements.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/general/latest/gr/wellarchitected.html

FAQ: https://aws.amazon.com/well-architected-tool/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/wellarchitected/latest/userguide/intro.html

User Guide: https://docs.aws.amazon.com/wellarchitected/latest/userguide/getting-started.html

API Reference: https://docs.aws.amazon.com/wellarchitected/latest/APIReference/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/well-architected-tool/pricing/

Cost Optimization

AWS Well-Architected Tool is completely free with no additional charges, providing significant cost optimization value as organizations only pay for underlying AWS resources used during assessments. The tool delivers cost optimization recommendations through its Cost Optimization pillar, helping identify opportunities like rightsizing instances, implementing Savings Plans, using Spot Instances, and eliminating idle resources:

- Integration with Cost Explorer and Trusted Advisor provides actionable cost reduction insights
- Organizations can track cost optimization improvements through milestone comparisons and quantify business value from optimization efforts
- The tool's assessment framework helps establish cost-aware cultures and implement financial management best practices across workloads

Savings Considerations

The primary cost savings from AWS Well-Architected Tool come from the optimization recommendations it provides rather than service costs, as the tool is free. Organizations achieve savings through improved architecture decisions, elimination of over-provisioned resources, implementation of appropriate pricing models like Reserved Instances or Savings Plans, and removal of unused resources identified through assessments:

- The tool's Cost Optimization pillar systematically identifies opportunities for expenditure reduction, demand management, and resource efficiency improvements
- Long-term savings result from establishing cost optimization practices, implementing automation for resource lifecycle management, and creating cost-aware organizational cultures
- Regular reviews using the tool help maintain optimal cost postures as workloads evolve and AWS service offerings expand

Service Name

Elastic Load Balancing (ELB)

Service Overview

Elastic Load Balancing (ELB) automatically distributes incoming application traffic across multiple targets, such as EC2 instances, containers, IP addresses, and Lambda functions, in one or more Availability Zones. It monitors the health of registered targets and routes traffic only to healthy targets. ELB scales load balancer capacity automatically in response to changes in incoming traffic, increasing application availability and fault tolerance by distributing workloads across multiple compute resources.

Key Use Cases

- Web application traffic distribution: Distributing HTTP/HTTPS traffic across multiple web servers for high availability and scalability
- Microservices architecture: Load balancing traffic to containerized applications and microservices with content-based routing
- TCP/UDP application load balancing: Distributing network traffic for database connections, gaming applications, and IoT workloads
- Auto Scaling integration: Automatically registering and deregistering instances as Auto Scaling groups scale up or down
- SSL/TLS termination: Offloading encryption and decryption tasks from backend servers to improve performance
- Multi-Availability Zone deployment: Ensuring fault tolerance by distributing traffic across multiple AZs
- Virtual appliance deployment: Using Gateway Load Balancers to deploy and scale third-party security appliances

Features

- **Application Load Balancer (ALB)**: Layer 7 load balancer with HTTP/HTTPS, gRPC support, content-based routing, and Lambda target integration
- **Network Load Balancer (NLB)**: Layer 4 load balancer with TCP/UDP/TLS support, static IP addresses, and extreme performance
- **Gateway Load Balancer**: Layer 3 gateway and layer 4 load balancing for deploying and scaling third-party virtual appliances
- **Health Checks**: Monitors target health and routes traffic only to healthy instances with configurable check parameters
- **Auto Scaling Integration**: Automatic registration and deregistration of instances with Auto Scaling groups

- **SSL/TLS Offloading**: Terminates SSL/TLS connections at the load balancer to reduce compute resource load
- **Cross-Zone Load Balancing**: Distributes traffic evenly across all enabled Availability Zones
- **Sticky Sessions**: Maintains session affinity to bind user sessions to specific backend instances
- **WebSocket Support**: Supports WebSocket connections for real-time applications
- **Content-Based Routing**: Routes requests based on URL path, host headers, HTTP methods, and query parameters

Value Proposition

Customers should choose Elastic Load Balancing for its automatic scaling capabilities, high availability across multiple Availability Zones, and seamless integration with other AWS services. ELB eliminates single points of failure, provides fault tolerance, and scales automatically to handle traffic growth without manual intervention:

- The service offers multiple load balancer types optimized for different use cases, from web applications to network protocols
- ELB reduces operational overhead by managing load balancer infrastructure, provides built-in security features like SSL/TLS termination, and offers pay-as-you-go pricing with no upfront costs or minimum commitments

Service Integration

ELB integrates deeply with core AWS services including Amazon EC2 for target instances, EC2 Auto Scaling for automatic instance management, Amazon VPC for network isolation, Route 53 for DNS routing, AWS Certificate Manager for SSL/TLS certificates, Amazon CloudWatch for monitoring and alerting, AWS WAF for web application firewall protection, Amazon ECS and EKS for containerized applications, AWS Lambda for serverless functions as targets, and AWS Global Accelerator for global traffic distribution. These integrations provide comprehensive application delivery solutions with minimal configuration complexity.

Onboarding and Offboarding

Customers get started by creating a load balancer through the AWS Management Console, AWS CLI, or APIs. The process involves selecting the load balancer type (Application, Network, or Gateway), configuring listeners and target groups, registering targets (EC2 instances, IP addresses, or Lambda functions), and setting up health checks:

- AWS provides step-by-step tutorials, getting started guides, and integration examples
- The setup can be automated using AWS CloudFormation or integrated into CI/CD pipelines

- Initial configuration typically takes minutes, with automatic scaling and health monitoring beginning immediately after deployment

Getting Started Guide: https://docs.aws.amazon.com/elasticloadbalancing/latest/userguide/what-is-load-balancing.html

Data Removal

Load balancer deletion removes the load balancer resource and stops traffic routing, but does not affect registered targets like EC2 instances. Customers must delete load balancers, target groups, and deregister targets manually. Access logs stored in S3 buckets must be deleted separately. CloudWatch metrics and logs are retained according to their configured retention periods. No customer data is stored within the load balancer service itself, simplifying the offboarding process.

Backup/Restore and Disaster Recovery

Elastic Load Balancing provides high availability through multi-AZ deployment but does not offer traditional backup capabilities as it is a stateless service. Configuration can be backed up using AWS Config or CloudFormation templates. For disaster recovery, load balancers can be recreated in different regions using Infrastructure as Code. The service supports cross-region failover when combined with Route 53 health checks and Global Accelerator for automated traffic shifting during outages.

Backup Capabilities

ELB does not have traditional backup capabilities as it is a stateless service that does not store customer data. Load balancer configurations can be captured using AWS Config for compliance and change tracking:

- Infrastructure as Code tools like CloudFormation, Terraform, or CDK can version and recreate load balancer configurations
- ELB itself does not integrate with AWS Backup service as there is no persistent data to backup

Disaster Recovery

ELB supports disaster recovery through multi-region deployment strategies rather than direct integration with AWS Elastic Disaster Recovery. Customers can deploy load balancers across multiple regions and use Route 53 for DNS failover:

- Load balancers automatically distribute traffic across healthy Availability Zones within a region
- For cross-region DR, Global Accelerator can provide automatic failover capabilities when combined with health checks and routing policies

Recovery Objectives

ELB helps meet aggressive RTO and RPO objectives through automatic failover capabilities within seconds when targets become unhealthy. Cross-zone load balancing ensures traffic redistribution occurs immediately when AZ failures occur. Multi-region deployments with Route 53 can achieve RTO of minutes for regional failures. Since ELB is stateless, RPO is effectively zero as no customer data is at risk. Health check frequency can be configured to detect failures within seconds.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/general/latest/gr/elb.html

FAQ: https://aws.amazon.com/elasticloadbalancing/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/elasticloadbalancing/latest/userguide/index.html

User Guide: https://docs.aws.amazon.com/elasticloadbalancing/latest/userguide/what-is-load-balancing.html

API Reference: https://docs.aws.amazon.com/elasticloadbalancing/latest/APIReference/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/elasticloadbalancing/pricing/

Cost Optimization

ELB offers several unique cost optimization strategies including Load Balancer Capacity Unit (LCU) reservations for predictable workloads, cross-zone load balancing optimization to reduce data transfer costs, and right-sizing load balancer types based on traffic patterns. Customers can optimize costs by consolidating multiple applications behind a single ALB using host-based or path-based routing, using Network Load Balancers for high-throughput TCP traffic to avoid ALB processing overhead, and leveraging AWS Free Tier offering 750 hours monthly for Classic Load Balancers. Monitoring CloudWatch metrics helps identify over-provisioned capacity and optimize target group configurations.

Savings Considerations

Main cost savings come from eliminating the need to provision and maintain dedicated load balancer infrastructure, paying only for actual usage through hourly charges and

processed data. Customers save on operational costs through automatic scaling, health management, and SSL certificate management:

- Consolidating multiple applications reduces the number of load balancers needed
- Using appropriate load balancer types prevents over-provisioning - Network Load Balancers for simple TCP routing cost less than Application Load Balancers
- Reserved LCU capacity provides up to 10% discounts for consistent workloads, and efficient target group design minimizes unnecessary health check costs and data processing charges

Service Name

AWS End User Messaging

Service Overview

AWS End User Messaging is a comprehensive messaging service that enables businesses to communicate with end users across multiple channels including SMS, MMS, voice messages, push notifications, and WhatsApp messaging. The service provides application-to-person (A2P) messaging capabilities with global scale, resiliency, and flexibility for delivering critical communications in web, mobile, or business applications. It consists of three main components: End User Messaging SMS for text and multimedia messaging, End User Messaging Push for mobile push notifications across various platforms, and End User Messaging Social for WhatsApp Business messaging integration.

Key Use Cases

- One-time password (OTP) authentication and login verification for secure user access
- Marketing campaigns and promotional messages with multimedia content and interactive features
- Critical notifications including appointment reminders, delivery status updates, and emergency alerts
- Customer engagement through multi-channel messaging workflows combining SMS, push, and WhatsApp
- Business communication workflows including customer support, order confirmations, and transactional updates

Features

- **SMS and MMS Messaging**: Global SMS and multimedia messaging with phone pools, sender IDs, and configuration sets
- **Voice Messaging**: Text-to-speech voice message delivery to phone numbers globally
- **Push Notifications**: Multi-platform push notifications supporting APNs, FCM, ADM, and Baidu Cloud Push
- **WhatsApp Business Integration**: Native WhatsApp messaging with rich media, interactive buttons, and template support
- **Phone Number Management**: Automated provisioning and management of origination identities including short codes, long codes, and toll-free numbers
- **SMS Protect**: Advanced security with artificial traffic filtering and country-based blocking configurations

- **Message Feedback Tracking**: Performance monitoring for conversion rates, click-through rates, and engagement metrics
- **Event Integration**: Integration with CloudWatch, SNS, and EventBridge for delivery receipts and notifications

Value Proposition

AWS End User Messaging provides a unified, enterprise-grade messaging platform that eliminates the complexity of managing multiple messaging vendors and technologies. The service offers global reach across 30+ AWS regions with built-in high availability and redundancy, automatic scaling to handle peak traffic volumes, and comprehensive security features including SMS Protect for fraud prevention:

- Organizations benefit from simplified integration through a single API for all messaging channels, reduced operational overhead with managed infrastructure, and cost optimization through pay-as-you-use pricing models
- The service includes advanced features like automated phone number registration, message feedback tracking, and seamless integration with other AWS services for comprehensive customer engagement workflows

Service Integration

AWS End User Messaging integrates deeply with core AWS services to provide comprehensive messaging workflows. Amazon CloudWatch provides monitoring, logging, and alerting for message delivery and performance metrics:

- Amazon SNS enables event-driven architectures and cross-service notifications
- AWS EventBridge facilitates real-time event routing and integration with third-party systems
- Amazon S3 stores media files for MMS messaging and message archives
- AWS Lambda enables serverless message processing and custom business logic integration
- Amazon DynamoDB supports contact management and opt-out list storage
- AWS IAM provides fine-grained access control and security policies
- The service also integrates with AWS Backup for data protection and AWS CloudFormation for infrastructure as code deployment across multiple accounts and regions

Onboarding and Offboarding

Getting started with AWS End User Messaging requires creating an AWS account and configuring the specific messaging channels needed. New accounts begin in a sandbox environment with spending limits and restricted capabilities that require service limit increase requests for production access:

- The onboarding process involves setting up phone pools or origination identities, configuring event destinations for delivery receipts, creating protect configurations for security, and implementing opt-out list management
- Customers can access the service through the AWS Management Console, AWS CLI, or AWS SDKs with comprehensive documentation and tutorials provided
- The service includes SMS simulators for testing without incurring costs and automated onboarding progress alerts via EventBridge to track phone number registration status

Getting Started Guide: https://docs.aws.amazon.com/sms-voice/latest/userguide/getting-started.html

Data Removal

Data removal and offboarding from AWS End User Messaging involves deleting messaging resources through the console or API, including phone numbers, sender IDs, configuration sets, and opt-out lists. Customers maintain control over their messaging data and can export contact lists and message logs before termination. The service automatically handles cleanup of temporary resources and staging areas. Message history and delivery receipts stored in integrated services like CloudWatch Logs or S3 require separate deletion according to configured retention policies. Account closure follows standard AWS procedures with final billing reconciliation.

Backup/Restore and Disaster Recovery

AWS End User Messaging provides built-in redundancy through multi-Availability Zone deployment within each region and native high availability architecture. The service supports cross-region failover strategies for resilient SMS delivery by configuring origination numbers in multiple regions. Message configurations, phone pools, and opt-out lists can be replicated across regions using Infrastructure as Code practices. Integration with AWS services enables comprehensive backup strategies including CloudWatch Logs retention, S3 storage policies, and EventBridge event archival for compliance and disaster recovery requirements.

Backup Capabilities

The service does not directly integrate with AWS Backup as messaging infrastructure is fully managed by AWS. However, customer configuration data, contact lists, and message templates can be backed up using AWS services like S3 for storage and DynamoDB for structured data:

- Message delivery logs and events are preserved in CloudWatch Logs according to configured retention policies
- Customers can export opt-out lists, phone number configurations, and other account-specific data for backup and compliance purposes

Disaster Recovery

AWS End User Messaging supports disaster recovery through multi-region architecture and automated failover capabilities. The service does not integrate with AWS Elastic Disaster Recovery as it provides native resilience through geographic distribution:

- Organizations can implement cross-region redundancy by provisioning origination identities in multiple regions and using DNS or application-level routing for failover
- The managed service architecture ensures automatic recovery from infrastructure failures within regions

Recovery Objectives

AWS End User Messaging helps customers achieve stringent RPO and RTO objectives through its globally distributed, highly available architecture. The service provides near-zero RPO for configuration data through real-time replication across Availability Zones and supports RTO targets under 1 hour through automated failover mechanisms. Multi-region deployment strategies enable customers to achieve sub-hour recovery times by implementing phone pool redundancy and automated traffic routing between regions during service disruptions.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/general/latest/gr/end-user-messaging.html

FAQ: https://docs.aws.amazon.com/sms-voice/latest/userguide/registrations-sg-faq.html

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/sms-voice/latest/userguide/what-is-sms-mms.html

User Guide: https://docs.aws.amazon.com/sms-voice/latest/userguide/index.html

API Reference: https://docs.aws.amazon.com/pinpoint/latest/apireference_smsvoicev2/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/end-user-messaging/pricing/

Cost Optimization

AWS End User Messaging offers unique cost optimization through its pay-per-message pricing model with no upfront costs or monthly minimums. Organizations can optimize costs by using phone pools for automatic origination identity selection, implementing message feedback tracking to improve conversion rates and reduce waste, and leveraging SMS Protect to filter artificial traffic and prevent fraud-related charges:

- Geographic routing strategies can reduce costs by using local numbers in each region, while message template optimization and character count management help minimize message segmentation costs
- The service's integration with AWS Cost Management tools enables detailed cost analysis and budget alerts

Savings Considerations

Primary cost savings come from eliminating the need for multiple messaging vendors and reducing operational overhead through managed infrastructure. Organizations save on integration costs through unified APIs across all messaging channels and automated phone number management:

- The pay-as-you-use model eliminates waste from over-provisioning, while message feedback tracking helps optimize campaign performance and ROI
- Additional savings result from reduced development time through comprehensive SDKs and documentation, elimination of infrastructure maintenance costs, and the ability to scale automatically without capacity planning
- Integration with AWS Free Tier provides cost-free testing and development environments

Service Name

AWS FreeRTOS

Service Overview

FreeRTOS is a market-leading, open-source real-time operating system (RTOS) for microcontrollers and small microprocessors, developed in partnership with leading chip companies and distributed under the MIT open source license. It includes the FreeRTOS kernel and a growing set of libraries suitable for connectivity, security, and over-the-air updates. FreeRTOS makes it easy to program, deploy, secure, connect, and manage small, low-power edge devices, extending the FreeRTOS kernel with software libraries that enable secure connections to AWS Cloud services or more powerful edge devices running AWS IoT Greengrass. The service emphasizes reliability, ease of use, and provides comprehensive IoT reference integrations for various industry sectors.

Key Use Cases

- Smart meters: Monitor electricity usage in real time and enable efficient load balancing for utilities
- Oil pump sensors: Monitor output on wells and process data locally via AWS IoT Greengrass Core for real-time analytics
- Industrial IoT devices: Build connected security systems, fitness trackers, and sensor networks for commercial applications
- Consumer appliances: Enable home automation, voice control, and connected home product development at scale
- Edge computing: Process data locally on microcontrollers while maintaining cloud connectivity for management and analytics

Features

- **FreeRTOS Kernel**: Market-leading real-time operating system kernel supporting 40+ architectures including RISC-V and ARMv8-M
- **IoT Connectivity Libraries**: Libraries for MQTT, HTTP, TCP/IP, Wi-Fi, Bluetooth Low Energy, and cellular connectivity
- **Security Libraries**: TLS v1.2, data encryption, key management, code signing, and secure communications
- **AWS IoT Integration**: Native integration with AWS IoT Core, Device Shadows, Device Defender, and Device Management
- **Over-the-Air Updates**: Secure firmware updates and device management capabilities with code signing
- **Long Term Support**: LTS releases with security updates and critical bug fixes for at least two years

- **Extended Maintenance Plan**: Up to 10 additional years of security patches and critical bug fixes beyond LTS

Value Proposition

FreeRTOS provides a unique combination of enterprise-grade reliability with open-source flexibility, backed by AWS's cloud infrastructure expertise. It offers the only RTOS with native AWS IoT service integration, eliminating complex integration work:

- Key differentiators include: proven stability from 18 years of development with leading chip partners, comprehensive security libraries including TLS and code signing, seamless AWS IoT connectivity with Device Shadows and OTA updates, extensive hardware support through the AWS Partner Device Catalog, and flexible licensing under MIT without vendor lock-in
- The service reduces development time and costs while providing enterprise-grade security and scalability for IoT deployments

Service Integration

FreeRTOS integrates natively with core AWS IoT services to provide comprehensive IoT solutions. Primary integrations include AWS IoT Core for device connectivity and messaging, AWS IoT Device Shadows for bi-directional device state management, AWS IoT Device Management for fleet provisioning and OTA updates, AWS IoT Device Defender for security monitoring and threat detection, and AWS IoT Greengrass for edge computing capabilities:

- Additional integrations span AWS Lambda for serverless functions, Amazon S3 for data storage, Amazon DynamoDB for device metadata, AWS Identity and Access Management for security policies, and AWS CloudWatch for monitoring and logging
- These integrations enable end-to-end IoT solutions from device to cloud analytics

Onboarding and Offboarding

Getting started with FreeRTOS involves downloading the latest release from FreeRTOS.org, selecting qualified hardware from the AWS Partner Device Catalog, and following board-specific getting started guides. Developers can explore functionality using the Windows Device Simulator without purchasing hardware:

- The process includes setting up the development environment with supported IDEs (Eclipse, Keil µVision), configuring AWS IoT credentials and certificates, building and flashing demo applications, and validating connectivity with AWS IoT Device Tester
- Comprehensive documentation, tutorials, and community forums provide step-by-step guidance
- For production deployment, customers integrate FreeRTOS libraries into custom applications and use AWS IoT services for device management and monitoring

Getting Started Guide: https://docs.aws.amazon.com/freertos/latest/userguide/freertos-getting-started.html

Data Removal

FreeRTOS offboarding involves removing device certificates from AWS IoT Device Registry, deleting device shadows and associated metadata, and deprovisioning devices from AWS IoT Device Management. Since FreeRTOS runs locally on devices, customers control all device-side data removal by reflashing firmware or resetting devices. Cloud-side data in AWS IoT services follows standard AWS data deletion practices through service APIs or console interfaces. No specific FreeRTOS offboarding process exists as the software operates independently on customer hardware.

Backup/Restore and Disaster Recovery

FreeRTOS operates as embedded software on customer hardware without built-in backup capabilities. Device state persistence relies on AWS IoT Device Shadows for cloud-based state storage and OTA update mechanisms for firmware recovery. Customers implement application-level backup strategies using AWS IoT services, storing critical data in Amazon S3 or DynamoDB. Disaster recovery involves reflashing firmware, restoring device configurations, and re-establishing AWS IoT connectivity. The distributed nature of edge devices provides inherent resilience against single points of failure.

Backup Capabilities

FreeRTOS does not provide traditional backup capabilities as it operates on resource-constrained microcontrollers. Device state backup occurs through AWS IoT Device Shadows which maintain device state in the cloud:

- Firmware backup involves storing images in device flash memory for OTA updates
- AWS services used with FreeRTOS (IoT Core, S3, DynamoDB) provide their own backup mechanisms
- Customers must implement application-level data persistence strategies suitable for their specific hardware constraints and use cases

Disaster Recovery

FreeRTOS supports disaster recovery through its distributed architecture and AWS cloud integrations. Device-level recovery involves OTA updates to restore firmware, Device Shadow synchronization to restore state, and automatic reconnection to AWS IoT services:

- Cloud-side recovery leverages standard AWS disaster recovery capabilities across IoT Core, Device Management, and storage services
- The decentralized nature of FreeRTOS deployments provides natural resilience, with individual device failures not affecting overall system operation

Recovery Objectives

FreeRTOS enables rapid recovery through OTA updates which can restore devices within minutes to hours depending on firmware size and network conditions. Device Shadow synchronization provides near real-time state recovery with RPO measured in seconds. Cloud service integration allows leveraging AWS disaster recovery capabilities with customizable RTO/RPO objectives. The distributed edge architecture ensures individual device failures don't impact overall system availability, supporting aggressive recovery objectives for critical IoT applications.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/general/latest/gr/amazon-freertos.html

FAQ: https://aws.amazon.com/freertos/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/freertos/latest/userguide/what-is-freertos.html

User Guide: https://docs.aws.amazon.com/freertos/latest/userguide/what-is-freertos.html

API Reference: https://docs.aws.amazon.com/freertos/latest/lib-ref/html1/annotated.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/freertos/pricing/

Cost Optimization

FreeRTOS offers exceptional cost optimization through its free MIT license with no runtime fees or royalties. The Extended Maintenance Plan provides predictable costs at $40,000 annually for single products or $90,000 for multiple products, enabling long-term budget planning:

- Customers can reduce infrastructure costs using low-cost microcontroller hardware and serverless AWS IoT services that scale automatically
- The efficient resource utilization minimizes hardware requirements, while local processing capabilities reduce data transfer costs
- Integration with AWS services enables pay-as-you-use models rather than fixed licensing fees, making it cost-effective for both small and large-scale deployments

Savings Considerations

Primary cost savings stem from eliminating RTOS licensing fees through the open-source MIT license, reducing development time with pre-built AWS integrations, and minimizing hardware costs through efficient resource utilization. Edge computing capabilities reduce cloud processing and data transfer costs by handling operations locally:

- The extensive qualified hardware ecosystem prevents vendor lock-in and enables competitive sourcing
- Long-term savings include reduced maintenance overhead through LTS support, predictable Extended Maintenance Plan pricing, and automatic scaling of AWS IoT services
- OTA update capabilities reduce field service costs, while comprehensive security libraries prevent costly security breaches and compliance issues

Service Name

Amazon Comprehend Medical

Service Overview

Amazon Comprehend Medical is a HIPAA-eligible natural language processing (NLP) service that uses machine learning to extract medical insights from unstructured clinical text such as physician notes, discharge summaries, test results, and pathology reports. It identifies and extracts medical entities like conditions, medications, dosages, tests, treatments, and procedures, while also detecting Protected Health Information (PHI). The service links extracted entities to standardized medical ontologies including ICD-10-CM, RxNorm, and SNOMED CT codes to enable structured data analysis for healthcare and life sciences applications.

Key Use Cases

- Patient case management and outcome tracking by extracting structured data from clinical notes for improved care coordination
- Clinical research and trial management by analyzing large volumes of medical text to identify patient cohorts and treatment patterns
- Medical billing and healthcare revenue cycle management through automated coding and claim validation using ontology linking
- Population health analytics and management by processing clinical documents to identify health trends and risk factors
- Pharmacovigilance and drug safety monitoring by extracting medication information and adverse events from clinical text
- Insurance claim processing acceleration by automatically identifying medical conditions and treatments from clinical documentation

Features

- **Medical Named Entity and Relationship Extraction (NERe)**: Identifies and extracts medical entities and their relationships from unstructured text, including medical conditions, medications, dosages, tests, treatments, and procedures with confidence scores
- **Protected Health Information (PHI) Detection**: Identifies and helps redact sensitive patient information to ensure compliance with data privacy regulations like HIPAA
- **Medical Ontology Linking**: Links extracted medical entities to standardized medical knowledge bases including ICD-10-CM, RxNorm, and SNOMED CT for consistent coding and interoperability

- **Batch and Real-time Processing**: Supports both synchronous operations for individual documents and asynchronous batch operations for collections of documents stored in Amazon S3
- **Confidence Scoring**: Provides confidence scores for each detected entity and linked concept to enable appropriate thresholding and human review workflows

Value Proposition

Amazon Comprehend Medical offers healthcare organizations a fully managed, HIPAA-eligible solution that eliminates the need to build, train, or deploy custom NLP models. It provides industry-leading accuracy in extracting medical information from clinical text while ensuring data security and compliance:

- The service integrates seamlessly with other AWS services for comprehensive data processing workflows and offers pay-as-you-use pricing with no upfront commitments
- Its pre-trained models are specifically designed for healthcare terminology and contexts, reducing development time and improving accuracy compared to generic NLP solutions
- The service scales automatically to handle varying workloads and provides standardized medical coding through ontology linking

Service Integration

Amazon Comprehend Medical integrates natively with Amazon S3 for batch processing of clinical documents stored in buckets, with input and output file management handled automatically. It works with Amazon SageMaker for building enhanced ML workflows and custom model development:

- Integration with Amazon Kinesis enables real-time data processing pipelines for streaming clinical text analysis
- The service connects with AWS HealthLake for centralized healthcare data storage and analysis, Amazon Athena for SQL-based querying of processed results, and AWS Lambda for serverless automation of text analysis workflows
- CloudWatch integration provides monitoring and logging capabilities, while IAM ensures secure access control and role-based permissions

Onboarding and Offboarding

Customers can get started by creating an AWS account and setting up an IAM user with appropriate permissions for Amazon Comprehend Medical. The service can be accessed through the AWS Management Console, AWS CLI, or AWS SDKs for various programming languages:

- Initial setup involves configuring AWS CLI credentials and testing the service with sample clinical text using the DetectEntitiesV2 API operation
- For batch processing, customers need to create S3 buckets for input and output files and configure IAM roles with proper bucket access permissions
- The service includes a free tier offering 8.5 million characters of text processing for the first month, allowing customers to evaluate capabilities before full implementation

Getting Started Guide: https://docs.aws.amazon.com/comprehend-medical/latest/dev/comprehendmedical-gettingstarted.html

Data Removal

Amazon Comprehend Medical processes data transiently and does not store customer content beyond the processing duration. For offboarding, customers should stop all active batch analysis jobs, delete any stored results from S3 output buckets, and revoke IAM permissions. The service does not retain processed text or extracted entities after analysis completion. Customers maintain full control over their data throughout the process and can delete all associated resources and outputs at any time through standard AWS resource management procedures.

Backup/Restore and Disaster Recovery

Amazon Comprehend Medical is a fully managed service where AWS handles infrastructure resilience and disaster recovery for the service itself. Customers are responsible for backing up their input data stored in S3 buckets and any processed outputs. The service operates across multiple Availability Zones within regions for high availability. For customer data protection, organizations should implement S3 Cross-Region Replication for input documents and results, and use S3 versioning to protect against accidental deletion or modification.

Backup Capabilities

Amazon Comprehend Medical does not provide built-in backup capabilities as it is a processing service that does not persistently store customer data. The service processes text transiently and returns results immediately for synchronous operations or stores results in customer-specified S3 buckets for batch operations:

- Customers must implement their own backup strategies for input data and results using S3 features like Cross-Region Replication, versioning, and lifecycle policies
- AWS Backup can be used to backup S3 buckets containing clinical documents and analysis results

Disaster Recovery

Amazon Comprehend Medical operates as a regional service with built-in high availability across multiple Availability Zones. The service itself does not integrate directly with AWS Elastic Disaster Recovery as it is a processing service without persistent state:

- For disaster recovery, customers should implement cross-region backup strategies for their input data and results using S3 Cross-Region Replication
- Organizations can set up processing workflows in multiple regions and use Route 53 for failover routing to ensure continued access to NLP capabilities during regional outages

Recovery Objectives

Amazon Comprehend Medical helps customers achieve low Recovery Point Objectives (RPO) by processing data without persistent storage, eliminating data loss risks during service operations. For Recovery Time Objectives (RTO), the service's multi-AZ architecture provides rapid failover within regions. Customers can achieve sub-hour RTO by implementing multi-region architectures with pre-configured processing workflows and using automated failover mechanisms. The service's API-based architecture enables quick restoration of processing capabilities in alternate regions when combined with proper data replication strategies.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/comprehend-medical/latest/dev/comprehendmedical-quotas.html

FAQ: https://aws.amazon.com/comprehend/medical/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/comprehend-medical/latest/dev/comprehendmedical-welcome.html

User Guide: https://docs.aws.amazon.com/comprehend-medical/latest/dev/comprehendmedical-welcome.html

API Reference: https://docs.aws.amazon.com/comprehend-medical/latest/api/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/comprehend/medical/pricing/

Cost Optimization

Amazon Comprehend Medical offers several cost optimization approaches including careful selection between synchronous and batch processing based on use case requirements, as batch processing typically offers better pricing for large volumes. Organizations can optimize costs by implementing confidence threshold filtering to reduce processing of low-value text segments and by using the free tier of 8.5 million characters in the first month for evaluation and testing:

- Efficient text preprocessing to remove unnecessary content and strategic use of different API operations (entity detection vs
- ontology linking) based on specific requirements can significantly reduce processing costs
- Regional selection and data locality optimization can also impact overall costs

Savings Considerations

Key cost savings considerations include leveraging the substantial free tier offering for initial implementation and testing phases, implementing batch processing for high-volume scenarios to achieve better unit economics, and designing efficient text preprocessing workflows to eliminate unnecessary processing of non-clinical content. Organizations can achieve savings by carefully selecting which ontology linking operations to use based on specific business requirements, as different operations have varying costs:

- Strategic implementation of confidence-based filtering and human review workflows can reduce overall processing volumes while maintaining quality
- Additionally, integrating with other AWS services like Lambda and S3 lifecycle policies can optimize the overall solution cost structure

Service Name

Amazon Security Lake

Service Overview

Amazon Security Lake is a fully managed security data lake service that automatically centralizes security data from AWS environments, SaaS providers, on-premises, and third-party sources into a purpose-built data lake stored in your AWS account. The service analyzes security data to provide a comprehensive understanding of security posture across the entire organization while improving protection of workloads, applications, and data. Security Lake automates collection of security-related log and event data, manages data lifecycle with customizable retention and replication settings, and converts ingested data into Apache Parquet format using the Open Cybersecurity Schema Framework (OCSF) standard.

Key Use Cases

- Centralized security data analysis: Aggregate and analyze security logs from multiple AWS services, SaaS applications, and on-premises systems for comprehensive threat detection
- Incident response and forensics: Provide security teams with normalized data in OCSF format for faster investigation and response to security events across multi-cloud environments
- Security analytics and machine learning: Enable advanced analytics using Amazon Athena, Amazon QuickSight, or third-party tools to identify patterns and anomalies in security data

Features

- **Data Aggregation**: Creates purpose-built security data lake in customer account with automated collection from cloud, on-premises, and custom sources
- **OCSF Normalization**: Transforms data from various sources to Open Cybersecurity Schema Framework standard for consistent analysis
- **Multi-Account Management**: Supports multi-account and multi-Region data management with rollup regions for centralized analysis
- **Subscriber Access Control**: Provides multiple levels of access for subscribers with data access and query access options
- **Lifecycle Management**: Manages data lifecycle with customizable retention and replication settings using S3 storage classes

Value Proposition

Amazon Security Lake provides unique value by offering a fully managed security data lake that eliminates the complexity of building custom security analytics infrastructure. Unlike traditional SIEM solutions, Security Lake stores data in customer-owned S3 buckets while providing automatic normalization to OCSF standard, enabling seamless integration with existing security tools:

- The service reduces operational overhead through automated data collection, transformation, and lifecycle management while maintaining data ownership
- Key differentiators include native AWS service integration, support for multi-account organizations, and ability to work with both AWS and third-party security analytics tools without vendor lock-in

Service Integration

Security Lake deeply integrates with core AWS services for comprehensive functionality. Primary integrations include Amazon S3 for data storage, AWS Glue for data cataloging and crawling, AWS Lake Formation for access control, and Amazon EventBridge for event notifications:

- The service collects data from AWS CloudTrail, Amazon VPC Flow Logs, AWS Security Hub, Amazon EKS, Route 53 resolver logs, and AWS WAF
- Subscriber services include Amazon Athena for querying, Amazon OpenSearch Service for analytics, Amazon SageMaker for ML insights, and Amazon Detective for investigations
- Additional integrations with AWS IAM, AWS KMS, Amazon SQS, and AWS Lambda provide security, encryption, messaging, and serverless processing capabilities

Onboarding and Offboarding

Customers can get started with Security Lake through the AWS Management Console with streamlined setup that automatically creates necessary IAM roles, or programmatically using APIs/CLI requiring manual IAM role creation. Prerequisites include an AWS account and administrative user access:

- The service offers a 15-day free trial in supported regions
- Onboarding involves enabling Security Lake in desired regions, configuring data sources from AWS services or custom sources, and setting up subscribers for data consumption
- For organizations, customers can use AWS Organizations integration with delegated administrator capabilities
- The setup process automatically provisions S3 buckets, Glue databases, and Lake Formation tables required for operation

Getting Started Guide: https://docs.aws.amazon.com/security-lake/latest/userguide/getting-started.html

Data Removal

Offboarding involves disabling Security Lake through console, API, or CLI, which stops log collection while retaining existing settings, resources, and data. For custom sources, integrations must be disabled outside Security Lake console. Data removal requires manual deletion of S3 buckets and associated resources. When re-enabling after disabling, a new S3 bucket is created requiring manual data migration from old bucket using S3 Sync command. Complete cleanup includes deleting Glue databases, Lake Formation resources, and IAM roles created during setup.

Backup/Restore and Disaster Recovery

Security Lake leverages AWS global infrastructure with high availability across Regions and Availability Zones. The service uses Amazon S3 for data storage with built-in durability and resilience features including lifecycle configurations, versioning, and multiple storage classes. S3 provides automatic backup capabilities with cross-region replication options. However, Security Lake itself does not provide specific backup/restore functionality beyond standard S3 capabilities. Data resiliency depends on underlying S3 features and customer-configured lifecycle policies for archiving and retention management.

Backup Capabilities

Security Lake does not have native backup capabilities but relies on Amazon S3 underlying infrastructure for data durability and availability. The service does not integrate with AWS Backup service:

- Data protection is achieved through S3 features including versioning, lifecycle management, and storage class transitions
- Customers can configure S3 Cross-Region Replication for additional data protection and implement custom backup strategies using S3 capabilities and lifecycle policies for long-term archival to Glacier storage classes

Disaster Recovery

Security Lake does not directly support AWS Elastic Disaster Recovery integration. Disaster recovery capabilities depend on underlying Amazon S3 infrastructure and multi-region deployment architecture:

- The service supports rollup regions for centralizing data from multiple contributing regions, providing some disaster recovery benefits
- Recovery relies on S3 cross-region replication, multi-region data distribution, and the inherent high availability of AWS infrastructure rather than specific disaster recovery tooling

Recovery Objectives

Security Lake helps meet RPO and RTO objectives through its multi-region architecture and S3-backed storage durability. The service enables low RPO through continuous data ingestion and S3's 99.999999999% durability. RTO improvements come from rollup region capabilities allowing quick access to centralized data and the ability to query data using multiple analytics tools like Athena. However, specific RPO/RTO metrics depend on customer configuration of data replication, retention policies, and subscriber access patterns rather than service guarantees.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/general/latest/gr/security-lake.html

FAQ: https://aws.amazon.com/security-lake/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/security-lake/latest/userguide/what-is-security-lake.html

User Guide: https://docs.aws.amazon.com/security-lake/latest/userguide/

API Reference: https://docs.aws.amazon.com/security-lake/latest/APIReference/

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/security-lake/pricing/

Cost Optimization

Security Lake offers several unique cost optimization options including a 15-day free trial, data lifecycle management with S3 storage class transitions to reduce long-term storage costs, and no charges for bringing third-party or custom data to the service. Customers can optimize costs by configuring appropriate data retention periods, leveraging S3 Intelligent Tiering for automatic cost optimization, and using lifecycle policies to transition data to cheaper storage classes like Glacier:

- The service pricing is based on data ingestion and normalization volumes, allowing customers to control costs by managing data source selection and retention policies
- Regional data processing helps minimize data transfer costs

Savings Considerations

Main cost savings come from eliminating the need to build and maintain custom security data infrastructure, reducing operational overhead through managed service benefits, and avoiding vendor lock-in with data stored in customer-owned S3 buckets. The service eliminates costs associated with separate SIEM licensing and infrastructure while providing automatic data normalization and transformation:

- Customers save on data engineering resources typically required for security data pipeline management
- The OCSF standardization reduces integration costs with multiple security tools
- Usage-based pricing model allows customers to pay only for data processed, and the ability to query data using existing AWS analytics services like Athena provides additional cost efficiencies compared to proprietary analytics platforms

Service Name

Amazon WorkMail

Service Overview

Amazon WorkMail is a secure, managed business email and calendaring service with support for existing desktop and mobile email clients. Users can access their email, contacts, and calendars using Microsoft Outlook, their browser, or native iOS and Android email applications. The service integrates with corporate directories and allows administrators to control data encryption keys and storage locations. Amazon WorkMail provides enterprise-grade security features including TLS encryption, multi-factor authentication, and data loss prevention while supporting Exchange ActiveSync protocol for mobile devices.

Key Use Cases

- Enterprise email management: Replace on-premises email servers with a fully managed cloud-based solution for business email and calendaring
- Microsoft Exchange migration: Migrate existing Exchange Server deployments to the cloud while maintaining interoperability with on-premises infrastructure
- Hybrid email environments: Enable bi-directional calendar sharing and unified global address books between Amazon WorkMail and on-premises Microsoft Exchange

Features

- **Email and Calendar Management**: Comprehensive email, contacts, and calendar functionality with 50 GB mailbox storage per user
- **Client Compatibility**: Native support for Microsoft Outlook, web browsers, mobile devices, and IMAP clients
- **Directory Integration**: Integration with AWS Directory Service, Microsoft Active Directory, and IAM Identity Center
- **Security Controls**: Data encryption at rest and in transit, mobile device policies, and remote wipe capabilities
- **Email Flow Rules**: Lambda-based email processing for filtering, content modification, and routing
- **Exchange Interoperability**: Bi-directional calendar sharing and unified address books with Microsoft Exchange
- **Mobile Device Management**: IT policy controls for mobile device security features and behavior
- **Email Journaling**: Compliance-ready email archiving and search capabilities for regulatory requirements

Value Proposition

Amazon WorkMail eliminates the need to manage on-premises email infrastructure while providing enterprise-grade security and compliance features. The service offers seamless integration with existing Microsoft environments through Exchange interoperability and Active Directory integration:

- Organizations benefit from pay-per-user pricing with no upfront costs, automatic scaling, and AWS-managed security updates
- The service provides strong data protection with customer-controlled encryption keys and data location control, making it ideal for organizations with strict security and compliance requirements

Service Integration

Amazon WorkMail deeply integrates with AWS Directory Service for user authentication and management, Amazon Simple Email Service (SES) for outbound email delivery, and AWS Key Management Service (KMS) for encryption key management. The service leverages AWS Lambda for email flow rules and content processing, Amazon CloudWatch for monitoring and logging, AWS CloudTrail for API audit logging, and Amazon S3 for event data storage. Integration with AWS IAM provides fine-grained access control, while Amazon VPC connectivity enables secure communication with on-premises resources.

Onboarding and Offboarding

Getting started involves creating an organization through the Amazon WorkMail console, adding an email domain, and setting up users either through new directory creation or integration with existing directories. Organizations can begin with a 30-day free trial supporting up to 25 users:

- The setup process includes DNS configuration for custom domains, user creation or enabling existing directory users, and configuring email clients
- Optional migration tools support transitioning from existing email solutions including Microsoft Exchange Server with full interoperability features

Getting Started Guide:
https://docs.aws.amazon.com/workmail/latest/adminguide/howto-start.html

Data Removal

Organization deletion is an irreversible process that permanently removes all mailbox data and associated resources. The process requires contacting AWS Support or using the DeleteOrganization API operation. Before deletion, organizations must remove any invalid Amazon SES rule sets and ensure Simple AD directories have no other AWS applications enabled. The deletion process includes options to retain or delete the associated directory and IAM Identity Center applications.

Backup/Restore and Disaster Recovery

Amazon WorkMail provides built-in resilience through AWS global infrastructure with multiple Availability Zones. Data protection includes automatic encryption at rest using unique keys per mailbox and AWS KMS integration. The service leverages AWS regional redundancy for high availability but does not provide explicit backup and restore capabilities for individual mailboxes or point-in-time recovery features.

Backup Capabilities

Amazon WorkMail does not integrate with AWS Backup service and does not provide traditional backup capabilities for mailbox data. The service relies on AWS infrastructure resilience and data replication across Availability Zones for data protection. Organizations requiring specific backup capabilities must implement third-party solutions or use email migration tools for data export.

Disaster Recovery

Amazon WorkMail does not integrate with AWS Elastic Disaster Recovery service. The service provides resilience through AWS multi-AZ infrastructure design with automatic failover capabilities built into the platform. Recovery relies on AWS regional infrastructure redundancy rather than explicit disaster recovery orchestration tools or services.

Recovery Objectives

Amazon WorkMail helps meet recovery objectives through AWS infrastructure built-in high availability and fault tolerance across multiple Availability Zones. The service provides automatic failover capabilities and leverages AWS global infrastructure for resilience. However, specific RPO and RTO metrics are not published, and recovery capabilities depend on AWS regional infrastructure rather than service-specific recovery features.

Service Constraints

Service Quotas:
https://docs.aws.amazon.com/workmail/latest/adminguide/workmail_limits.html

FAQ: https://aws.amazon.com/workmail/faqs/

Technical Requirements

Service Documentation:
https://docs.aws.amazon.com/workmail/latest/adminguide/what_is.html

User Guide: https://docs.aws.amazon.com/workmail/latest/userguide/what_is.html

API Reference:
https://docs.aws.amazon.com/workmail/latest/APIReference/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/workmail/pricing/

Cost Optimization

Amazon WorkMail offers straightforward per-user monthly pricing at $4.00 per user including 50 GB storage with no upfront fees or long-term commitments. Cost optimization focuses on user lifecycle management with prorated billing for new users and full-month charges for terminated users:

- Organizations can minimize costs by carefully managing active user counts, utilizing the 30-day free trial for up to 25 users, and avoiding unnecessary journaling features which incur additional charges
- The service eliminates infrastructure costs associated with on-premises email servers

Savings Considerations

Primary cost savings come from eliminating on-premises email infrastructure maintenance, hardware refresh cycles, and dedicated IT staff for email system management. The pay-per-user model eliminates over-provisioning costs common with traditional licensing models:

- Organizations save on security management costs through built-in enterprise-grade security features, automatic updates, and AWS-managed infrastructure
- Migration from Exchange Server reduces licensing costs, and the service scales automatically without capacity planning investments
- Operational savings include reduced downtime costs and simplified disaster recovery compared to self-managed solutions

Service Name

AWS CodeCommit

Service Overview

AWS CodeCommit is a fully managed source control service hosted by Amazon Web Services that enables organizations to privately store and manage assets such as documents, source code, and binary files in the cloud. CodeCommit eliminates the need to manage your own source control system or worry about scaling infrastructure by providing secure, highly scalable private Git repositories. The service supports standard Git functionality and works seamlessly with existing Git-based tools, enabling teams to collaborate through features like pull requests, branching, merging, and code reviews while ensuring high availability, durability, and enterprise-grade security.

Key Use Cases

- Source code management: Store and version control application source code with full Git functionality for software development teams
- CI/CD pipeline integration: Serve as the source repository for automated build, test, and deployment workflows using AWS CodePipeline, CodeBuild, and CodeDeploy
- Collaborative development: Enable team collaboration through pull requests, code reviews, branching, and merging with fine-grained access control
- Asset management: Store and manage documents, binaries, and other development assets alongside source code in secure private repositories
- Enterprise compliance: Meet regulatory and compliance requirements with encryption, audit trails, and IAM-based access controls for secure code storage

Features

- **Private Git Repositories**: Unlimited private Git repositories with full Git functionality including branching, merging, and tagging
- **Pull Requests and Code Reviews**: Built-in pull request workflows with approval rules, code review capabilities, and team collaboration features
- **Encryption and Security**: Data encrypted in transit via HTTPS/SSH and at rest using AWS KMS with fine-grained IAM access controls
- **High Availability and Durability**: Repositories stored across multiple Availability Zones using Amazon S3 and DynamoDB for enterprise-grade reliability
- **Event Notifications and Triggers**: Repository event notifications via Amazon SNS and configurable triggers to invoke AWS Lambda functions
- **Cross-Account Access**: Support for cross-account repository sharing and federated access using IAM roles and policies

- **Multiple Connection Methods**: Access via HTTPS Git credentials, SSH keys, git-remote-codecommit, and AWS CLI credential helper

Value Proposition

Organizations choose AWS CodeCommit for its deep integration with AWS services and unique enterprise capabilities that differentiate it from other Git hosting solutions. CodeCommit provides seamless IAM integration for fine-grained access control, VPC endpoint support for private network access, and comprehensive AWS CloudTrail logging for audit and compliance requirements:

- The service eliminates infrastructure management overhead while offering enterprise-grade security with AWS KMS encryption and FIPS 140-2 compliance
- CodeCommit's native integration with AWS CodePipeline, CodeBuild, and CodeDeploy enables frictionless CI/CD workflows within the AWS ecosystem, making it ideal for organizations already invested in AWS services or those requiring strict compliance and security controls

Service Integration

CodeCommit integrates natively with core AWS DevOps and development services to enable complete CI/CD workflows. Primary integrations include AWS CodePipeline for automated build and deployment pipelines, AWS CodeBuild for continuous integration and testing, and AWS CodeDeploy for application deployments:

- The service leverages AWS IAM for access control and authentication, AWS KMS for encryption key management, and Amazon CloudWatch Events for monitoring and notifications
- CodeCommit also integrates with Amazon SNS for event notifications, AWS Lambda for custom triggers, AWS CloudTrail for audit logging, and Amazon CodeGuru Reviewer for automated code reviews
- Additional integrations include AWS Amplify for frontend deployments, AWS Cloud9 for cloud-based development environments, and AWS Elastic Beanstalk for web application deployments

Onboarding and Offboarding

Getting started with CodeCommit requires setting up AWS credentials and configuring local Git access through multiple supported methods. Customers begin by creating an AWS account and configuring IAM permissions, then choose from connection methods including HTTPS Git credentials (simplest), SSH keys, git-remote-codecommit (recommended for federated access), or AWS CLI credential helper:

- The setup process involves creating repositories through the AWS Management Console, AWS CLI, or APIs, then cloning them locally using standard Git commands

- CodeCommit provides comprehensive tutorials covering repository creation, local setup, making commits, creating branches, and configuring team access permissions
- The service offers multiple getting started guides depending on Git familiarity and connection preferences, ensuring smooth onboarding for both new and experienced Git users

Getting Started Guide:
https://docs.aws.amazon.com/codecommit/latest/userguide/getting-started.html

Data Removal

CodeCommit offboarding involves deleting repositories through the AWS Management Console, AWS CLI, or API operations. Before deletion, customers should backup repository data using standard Git clone commands or AWS Backup service snapshots. Once repositories are deleted, all associated data including commit history, branches, tags, and metadata are permanently removed from CodeCommit storage. Customers can export repository data by cloning to local systems or pushing to alternative Git hosting services before deletion. The service provides confirmation prompts and supports programmatic deletion for bulk operations during migration scenarios.

Backup/Restore and Disaster Recovery

CodeCommit provides built-in high availability and disaster recovery through automatic replication across multiple Availability Zones within a region. The service stores repository data redundantly in Amazon S3 and Amazon DynamoDB for durability and availability. While CodeCommit itself is highly durable, customers can create additional backups using standard Git operations like cloning repositories or by integrating with AWS Backup service for automated snapshot management. Repository data can be restored from local clones or backup snapshots, and the distributed nature of Git provides inherent backup capabilities through multiple repository copies.

Backup Capabilities

CodeCommit offers native backup capabilities through AWS Backup service integration, enabling automated repository snapshots and centralized backup management. Customers can create manual backups using standard Git clone operations to preserve complete repository history, branches, and tags:

- The service's architecture provides inherent backup through automatic replication across multiple Availability Zones
- Repository data can be exported and backed up to external systems using Git operations, and deleted repositories can be restored from backups if created before deletion

Disaster Recovery

CodeCommit supports disaster recovery through its multi-Availability Zone architecture and integration with AWS disaster recovery services. The service automatically replicates data across multiple AZs for high availability and fault tolerance:

- While CodeCommit doesn't directly integrate with AWS Elastic Disaster Recovery, customers can implement DR strategies using Git's distributed nature by maintaining repository mirrors across regions or external systems
- Repository data can be quickly restored from local clones or backup snapshots in disaster scenarios

Recovery Objectives

CodeCommit helps customers achieve low Recovery Point Objectives (RPO) and Recovery Time Objectives (RTO) through its highly available architecture and Git's distributed nature. The service's automatic multi-AZ replication provides near-zero RPO for committed changes, while the ability to quickly clone repositories from backup sources enables rapid recovery. Customers can maintain multiple repository copies across regions or external systems to further reduce RTO. The service's integration with AWS Backup enables point-in-time recovery capabilities for more granular RPO requirements.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/codecommit/latest/userguide/limits.html

FAQ: https://aws.amazon.com/codecommit/faqs/

Technical Requirements

Service Documentation:
https://docs.aws.amazon.com/codecommit/latest/userguide/welcome.html

User Guide: https://docs.aws.amazon.com/codecommit/latest/userguide/welcome.html

API Reference:
https://docs.aws.amazon.com/codecommit/latest/APIReference/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/codecommit/pricing/

Cost Optimization

CodeCommit offers unique cost optimization through its usage-based pricing model charging only for active users (those performing Git operations monthly) rather than total users or repositories. The service includes a permanent free tier for up to 5 active users per month, unlimited private repositories, and no charges for data transfer or Git operations beyond storage and request quotas:

- Cost optimization strategies include managing active user counts by consolidating access through service accounts, using the service's unlimited repository feature to avoid per-repository fees, and leveraging the free storage and request allowances
- Organizations can optimize costs by timing development activities and using batch operations to minimize active user months

Savings Considerations

CodeCommit provides significant cost savings compared to self-managed Git hosting solutions by eliminating infrastructure provisioning, maintenance, and scaling costs. The service's operational model reduces total cost of ownership through automatic backup, security patching, and high availability without additional charges:

- Organizations save on licensing fees for commercial Git hosting platforms while gaining enterprise features like encryption, compliance certifications, and AWS service integration
- The pay-per-active-user model ensures costs scale with actual usage rather than provisioned capacity
- Additional savings come from reduced administrative overhead, faster deployment cycles through AWS service integration, and elimination of third-party tool licensing for CI/CD workflows when combined with other AWS DevOps services

Service Name

AWS Security Hub

Service Overview

AWS Security Hub is a unified cloud security solution that prioritizes critical security issues and helps organizations respond to threats at scale. Security Hub automatically correlates and enriches security signals from multiple sources including posture management, vulnerability management (Amazon Inspector), sensitive data discovery (Amazon Macie), and threat detection (Amazon GuardDuty). Through intuitive visualizations and near real-time risk analytics, it transforms complex security signals into actionable insights that enable informed security decisions. Security Hub includes automated response workflows, integrated ticketing systems, and exposure findings to help remediate risks, improve team productivity, and minimize operational disruptions across cloud environments.

Key Use Cases

- Centralized security operations: Unify security management across multiple AWS accounts and services with consolidated visibility and standardized finding formats
- Security posture management: Continuously monitor cloud configurations against security best practices and regulatory compliance frameworks like CIS, PCI DSS, and NIST
- Threat detection and response: Correlate security findings from multiple services to identify exposures, potential attack paths, and automate incident response workflows

Features

- **Exposure Findings**: Correlates findings from Security Hub CSPM, Amazon Inspector, and other AWS services to detect exploitable vulnerabilities and misconfigurations
- **Attack Path Visualization**: Interactive visualization showing how potential attackers can access and control resources associated with exposure findings
- **Automated Correlation**: Automatically correlates and enriches security signals across multiple AWS services to identify related findings
- **Multi-Account Support**: Centralized management across AWS Organizations with cross-Region aggregation of findings
- **Security Standards Compliance**: Automated security checks against industry frameworks including AWS Foundational Security Best Practices, CIS, PCI DSS, and NIST
- **Automation Rules**: Automated response workflows and integration with ticketing systems like Jira Cloud and ServiceNow ITSM

- **OCSF Integration**: Generates and receives findings in Open Cybersecurity Schema Framework for standardized data exchange
- **Historical Trends**: Up to 1 year of historical data with customizable widgets for exposures, threats, resources, and security coverage

Value Proposition

AWS Security Hub provides unmatched value through its unified approach to cloud security operations, eliminating the need to manage multiple disparate security tools and consoles. Unlike standalone security solutions, Security Hub automatically correlates findings across Amazon GuardDuty, Amazon Inspector, Amazon Macie, and AWS Config to provide contextualized risk prioritization and exposure analysis:

- Its streamlined pricing model consolidates charges across integrated security services, reducing complexity and cost
- The service offers near real-time risk analytics with automated correlation capabilities that transform overwhelming security alerts into prioritized, actionable insights
- Security Hub's deep AWS integration enables automated remediation workflows and seamless multi-account management through AWS Organizations, providing comprehensive security visibility without additional operational overhead

Service Integration

AWS Security Hub deeply integrates with core AWS security services to provide comprehensive coverage. Primary integrations include Amazon GuardDuty for threat detection, Amazon Inspector for vulnerability management, Amazon Macie for sensitive data discovery, and AWS Security Hub CSPM for cloud security posture management:

- Security Hub also integrates with AWS Config for configuration monitoring, AWS Identity and Access Management (IAM) for access control, AWS CloudTrail for API logging, and AWS Lambda for automated response functions
- Additional integrations include Amazon EventBridge for event routing, Amazon CloudWatch for monitoring and metrics, AWS Systems Manager for automated remediation, and AWS Organizations for multi-account management
- Third-party integrations are supported through standardized APIs, enabling connections with SIEM tools, ticketing systems like Jira and ServiceNow, and other security solutions

Onboarding and Offboarding

Customers can get started with AWS Security Hub through a straightforward onboarding process that begins with enabling Security Hub in the AWS Console. For individual accounts, users simply enable Security Hub and choose to enable all capabilities or customize them according to their needs:

- For AWS Organizations, the management account designates a delegated administrator who enables Security Hub and configures member accounts
- Prerequisites include enabling AWS Config and ensuring proper IAM permissions
- The service automatically creates necessary service-linked roles and recorders upon enablement
- Security Hub offers a 30-day free trial period, and customers can enable security standards like AWS Foundational Security Best Practices, CIS benchmarks, or PCI DSS based on their compliance requirements
- Integration with other AWS security services like GuardDuty, Inspector, and Macie can be configured simultaneously for comprehensive coverage

Getting Started Guide:
https://docs.aws.amazon.com/securityhub/latest/userguide/securityhub-v2-enable.html

Data Removal

Security Hub offboarding involves disabling the service through the AWS Console or API, which removes access to the Security Hub dashboard and stops new finding ingestion. Existing findings are retained according to the service's 90-day retention policy, after which they are automatically deleted. Customers can manually archive findings before offboarding if needed. The service-linked roles and associated resources are cleaned up when Security Hub is disabled. For Organizations, the delegated administrator can disable Security Hub across all member accounts. Customers should ensure any automated workflows or integrations dependent on Security Hub are updated before offboarding to prevent service disruptions.

Backup/Restore and Disaster Recovery

AWS Security Hub operates as a managed service where AWS handles infrastructure resilience and data durability. Security Hub findings and configurations are automatically replicated across multiple Availability Zones within a region for high availability. The service provides cross-Region aggregation capabilities allowing findings to be viewed from a central aggregation region for disaster recovery scenarios. However, Security Hub itself is not a backup service and does not provide backup capabilities for customer data. Organizations must implement separate backup strategies for their underlying AWS resources that Security Hub monitors and configure cross-Region aggregation to maintain security visibility during regional outages.

Backup Capabilities

Security Hub does not provide backup capabilities and is not integrated with AWS Backup. As a security management service, Security Hub focuses on aggregating and analyzing security findings rather than backing up data:

- The service maintains findings for 90 days with automatic retention management

- Organizations requiring backup of security configurations or findings data should implement custom solutions using APIs to export findings data or leverage third-party integrations for long-term storage and analysis of security data

Disaster Recovery

Security Hub supports disaster recovery through its cross-Region aggregation feature, which allows findings from multiple linked regions to be aggregated in a central region. This capability ensures security visibility is maintained even if individual regions experience outages:

- However, Security Hub does not integrate with AWS Elastic Disaster Recovery as it is not a workload that requires disaster recovery orchestration
- The service's multi-AZ architecture and AWS-managed infrastructure provide inherent resilience, and cross-Region aggregation serves as the primary disaster recovery mechanism for maintaining security operations continuity

Recovery Objectives

Security Hub helps customers meet RPO and RTO objectives through its cross-Region aggregation capabilities and near real-time finding correlation. The service provides continuous monitoring with minimal data loss potential (low RPO) through automated finding ingestion from integrated security services. For RTO, Security Hub's cross-Region aggregation ensures security visibility can be quickly restored from alternative regions during outages. The service's automated correlation and exposure findings help prioritize critical security issues, enabling faster incident response times. However, specific RPO/RTO values depend on the underlying AWS services and customer's cross-Region configuration strategy.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/general/latest/gr/sechub.html

FAQ: https://aws.amazon.com/security-hub/features/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/securityhub/latest/userguide/index.html

User Guide: https://docs.aws.amazon.com/securityhub/latest/userguide/what-is-securityhub-v2.html

API Reference: https://docs.aws.amazon.com/securityhub/1.0/APIReference/Welcome.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/security-hub/pricing/

Cost Optimization

AWS Security Hub offers unique cost optimization through its streamlined pricing model that consolidates charges across integrated AWS security services. The service uses resource-based pricing anchored on Amazon EC2 instances, eliminating complex per-finding charges:

- Security Hub includes a 30-day free trial and free tiers for finding ingestion events (10,000 per month) and automation rule evaluations (1 million per month)
- The Cost Estimator feature helps organizations compare costs with and without Security Hub, estimate costs before enabling capabilities, and export estimates for budgeting
- Volume pricing reduces per-resource costs for larger deployments
- By centralizing security operations, Security Hub reduces operational overhead and eliminates the need for multiple specialized security tools, providing cost efficiency through unified management and reduced complexity

Savings Considerations

Security Hub delivers significant cost savings by consolidating multiple security tools into a unified solution, reducing licensing costs for disparate security products. The service eliminates operational overhead associated with managing multiple security consoles and tools, reducing staffing costs for security operations teams:

- Automated correlation and prioritization reduce time spent on manual security analysis and false positive investigation
- Cross-Region aggregation minimizes the need for regional security teams and infrastructure
- The streamlined pricing model with volume discounts provides predictable costs compared to pay-per-finding alternatives
- Integration with AWS native services reduces data transfer costs and eliminates the need for third-party security data aggregation solutions
- Organizations can achieve 30-70% cost reduction in security operations through automated workflows and reduced manual intervention requirements

Service Name

AWS Service Catalog

Service Overview

AWS Service Catalog enables organizations to create and manage catalogs of IT services that are approved for use on AWS. It allows IT administrators to centrally manage and distribute cloud resources as standardized products to end users through a self-service portal. The service provides governance controls, version management, and fine-grained access control while enabling users to quickly deploy approved AWS resources without requiring direct access to underlying services. Service Catalog supports CloudFormation templates, Terraform configurations, and AWS Marketplace products.

Key Use Cases

- Self-service provisioning: Enable end users to deploy approved AWS resources through a personalized portal without requiring extensive AWS knowledge or permissions
- Governance and compliance: Centrally manage and enforce organizational policies, constraints, and standardized configurations for cloud resource deployment
- Multi-account resource management: Share portfolios and products across multiple AWS accounts while maintaining centralized control and governance

Features

- **Product Management**: Create and manage IT service products using CloudFormation templates, Terraform configurations, or AWS Marketplace offerings
- **Portfolio Organization**: Organize products into portfolios with customizable access controls and configuration information
- **Version Control**: Manage multiple versions of products and automatically propagate updates across portfolios
- **Constraints and Governance**: Apply launch constraints, notification constraints, and template constraints to control resource deployment
- **AppRegistry**: Manage applications and their associated configurations, resources, and metadata
- **Service Actions**: Define custom actions that can be performed on provisioned products for operational tasks

Value Proposition

AWS Service Catalog provides unique value through standardization of cloud resources, enabling self-service discovery and launch of approved products, fine-grained access

control through IAM integration, and extensibility with version control. Organizations benefit from reduced operational overhead, consistent governance across multi-account environments, faster time-to-market for development teams, and improved compliance through pre-approved resource templates. The service eliminates the need for complex IAM permissions while maintaining security controls.

Service Integration

Service Catalog deeply integrates with AWS CloudFormation for infrastructure deployment, AWS IAM for access control and role-based permissions, AWS Organizations for multi-account governance, and AWS CloudFormation StackSets for cross-account deployments. It also supports HashiCorp Terraform for infrastructure as code, integrates with AWS Marketplace for third-party products, works with AWS Backup for data protection strategies, and connects with external systems like ServiceNow and Jira Service Desk through APIs.

Onboarding and Offboarding

Customers begin by completing prerequisites in Setting Up AWS Service Catalog, then can use the Getting Started Library templates or follow step-by-step tutorials. The process involves creating portfolios as an administrator, adding products using CloudFormation templates or Terraform configurations, setting constraints and permissions, and granting access to end users:

- End users access the Service Catalog console to browse and launch approved products
- The service includes comprehensive APIs, CLI support, and integration capabilities for programmatic management

Getting Started Guide:
https://docs.aws.amazon.com/servicecatalog/latest/adminguide/getstarted.html

Data Removal

Data removal involves terminating provisioned products using TerminateProvisionedProduct API, which deletes associated CloudFormation stacks and resources. Administrators can delete products, portfolios, and associated constraints through the console or APIs. All provisioned resources are managed through underlying AWS services, so standard AWS data deletion processes apply to the actual infrastructure components deployed through Service Catalog.

Backup/Restore and Disaster Recovery

Service Catalog relies on underlying AWS services for backup and disaster recovery. Provisioned products inherit backup capabilities from their underlying resources (EC2, RDS, etc.). Organizations can implement AWS Backup policies for Service Catalog-

deployed resources and use AWS Elastic Disaster Recovery for comprehensive disaster recovery strategies. The service metadata and configurations can be replicated across regions.

Backup Capabilities

Service Catalog itself does not provide direct backup capabilities but enables backup through underlying AWS services. Provisioned products can leverage AWS Backup, service-specific backup features, and cross-region replication. Organizations should implement backup strategies for both Service Catalog configurations and the actual resources deployed through products.

Disaster Recovery

Service Catalog supports disaster recovery through integration with AWS Elastic Disaster Recovery and cross-region deployment capabilities. Organizations can deploy portfolios and products across multiple regions, use CloudFormation StackSets for multi-region provisioning, and implement disaster recovery strategies for underlying resources deployed through Service Catalog products.

Recovery Objectives

Service Catalog helps meet RPO and RTO objectives through automated provisioning, standardized recovery procedures, and integration with AWS disaster recovery services. Organizations can pre-define disaster recovery products, automate failover processes through Service Actions, and leverage the speed of infrastructure-as-code deployment to minimize recovery times during disaster scenarios.

Service Constraints

Service Quotas:
https://docs.aws.amazon.com/servicecatalog/latest/adminguide/limits.html

FAQ: https://aws.amazon.com/servicecatalog/faqs/

Technical Requirements

Service Documentation:
https://docs.aws.amazon.com/servicecatalog/latest/adminguide/what-is_concepts.html

User Guide:
https://docs.aws.amazon.com/servicecatalog/latest/adminguide/introduction.html

API Reference:
https://docs.aws.amazon.com/servicecatalog/latest/dg/API_Reference.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/servicecatalog/pricing/

Cost Optimization

Service Catalog offers unique cost optimization through standardized resource templates that prevent resource sprawl, automated rightsizing through constraints, and centralized cost management. Organizations can implement budget controls through Service Catalog budgets, use tagging strategies for cost allocation, and leverage the free tier of 1,000 API calls per month. The pay-per-API-call pricing model with no upfront costs makes it cost-effective for organizations of all sizes.

Savings Considerations

Primary cost savings come from preventing unauthorized or oversized resource deployments through governance controls, reducing operational overhead through self-service provisioning, and eliminating duplicate infrastructure deployments. Organizations save on administrative costs through automated workflows, reduce training expenses by providing standardized deployment processes, and achieve better resource utilization through consistent templates and constraints. The service's integration with AWS Cost Management tools enables comprehensive cost tracking and optimization.

Service Name

AWS X-Ray

Service Overview

AWS X-Ray is a distributed tracing service that helps developers analyze and debug production applications, APIs, and services. It provides deep visibility into application performance by tracking requests as they flow through microservices architectures, identifying bottlenecks, errors, and latency issues. X-Ray generates service maps that visualize application components and their relationships, enabling developers to understand how services interact with AWS resources and external systems. The service supports both simple applications and complex microservices architectures, offering end-to-end request tracing capabilities.

Key Use Cases

- Performance optimization: Identifying latency bottlenecks and performance issues across distributed applications and microservices
- Error debugging: Root cause analysis of application errors and exceptions in production environments
- Application monitoring: Understanding request flow and dependencies between services for operational visibility
- Microservices analysis: Visualizing complex service relationships and interactions in distributed architectures
- Capacity planning: Analyzing application behavior to optimize resource allocation and scaling decisions

Features

- **End-to-end tracing**: Tracks requests across multiple services and components to provide complete visibility into application behavior
- **Service map generation**: Creates visual representations of application architecture showing service dependencies and relationships
- **Latency detection**: Measures server-side and client-side latency to identify performance bottlenecks
- **Error analysis**: Automatically highlights bugs and errors in application code by analyzing response codes and exceptions
- **Data annotation and filtering**: Allows adding annotations and metadata to traces for easier analysis and filtering
- **Sampling configuration**: Customizable sampling rules to control trace volume and costs without code modifications

- **Multi-language SDK support**: SDKs available for Java, Node.js, Python, Ruby, Go, and .NET
- **Query and analytics APIs**: Programmatic access to trace data for building custom analysis and visualization applications

Value Proposition

AWS X-Ray provides unique advantages as a fully managed distributed tracing service with deep AWS integration. Unlike third-party solutions, X-Ray seamlessly integrates with AWS services like Lambda, API Gateway, ECS, and CloudWatch without requiring complex setup:

- It offers automatic instrumentation for many AWS services, reducing implementation overhead
- The service provides cost-effective pay-as-you-go pricing with generous free tiers, making it accessible for applications of all sizes
- X-Ray's service maps and trace analytics enable faster root cause analysis compared to traditional logging approaches, significantly reducing mean time to resolution for production issues

Service Integration

AWS X-Ray integrates natively with numerous AWS services including AWS Lambda (automatic tracing), Amazon API Gateway (request tracing), Amazon ECS and EKS (containerized applications), AWS Elastic Beanstalk (automatic instrumentation), Elastic Load Balancing (request routing visibility), Amazon CloudWatch (metrics and alarms integration), Amazon SNS and SQS (message tracing), Amazon DynamoDB and RDS (database call tracing), and Amazon EventBridge (event-driven architectures). X-Ray also works with CloudWatch Application Signals, CloudWatch RUM for client-side monitoring, and supports AWS Distro for OpenTelemetry for vendor-neutral instrumentation.

Onboarding and Offboarding

Getting started with AWS X-Ray involves three main steps: instrument applications using X-Ray SDKs or AWS Distro for OpenTelemetry, configure the X-Ray daemon or CloudWatch agent to collect and relay trace data, and deploy the instrumented application to begin generating traces. For AWS Lambda, tracing can be enabled with a simple configuration change:

- Applications can be instrumented using language-specific SDKs or by leveraging automatic instrumentation available for many AWS services
- The X-Ray console provides immediate visualization of trace data once collection begins

Getting Started Guide: https://docs.aws.amazon.com/xray/latest/devguide/xray-gettingstarted.html

Data Removal

AWS X-Ray automatically deletes trace data after 30 days of retention, requiring no manual intervention for data removal. To offboard completely, customers should disable tracing in applications by removing X-Ray SDK instrumentation or disabling tracing configuration, stop the X-Ray daemon or CloudWatch agent, and delete custom sampling rules and groups through the console or API. All trace data will be automatically purged after the 30-day retention period.

Backup/Restore and Disaster Recovery

AWS X-Ray does not provide traditional backup and disaster recovery capabilities as it is designed as an operational monitoring service. Trace data is automatically retained for 30 days and then deleted. X-Ray itself is a highly available service that operates across multiple Availability Zones within each region. For disaster recovery planning, customers should consider X-Ray as part of their application monitoring stack rather than a data persistence service requiring backup.

Backup Capabilities

AWS X-Ray does not have built-in backup capabilities and does not integrate with AWS Backup. The service is designed for operational monitoring with automatic 30-day data retention and deletion:

- Trace data is considered ephemeral operational telemetry rather than persistent business data requiring backup
- Customers needing long-term trace analysis should export data using X-Ray APIs before the 30-day retention period expires

Disaster Recovery

AWS X-Ray does not integrate with AWS Elastic Disaster Recovery as it is an operational monitoring service rather than a business-critical application or data service. X-Ray itself provides high availability within each AWS region. For multi-region disaster recovery scenarios, customers can enable X-Ray tracing in multiple regions and use cross-account observability features to centralize monitoring across regions and accounts.

Recovery Objectives

AWS X-Ray indirectly supports RTO and RPO objectives by providing rapid root cause analysis capabilities that accelerate incident resolution. The service enables faster identification of performance issues and errors, reducing mean time to resolution (MTTR). While X-Ray itself doesn’t provide backup or recovery capabilities, its monitoring and debugging features help teams meet recovery objectives by minimizing downtime through faster problem diagnosis and resolution during incidents.

Service Constraints

Service Quotas: https://docs.aws.amazon.com/general/latest/gr/xray.html

FAQ: https://aws.amazon.com/xray/faqs/

Technical Requirements

Service Documentation: https://docs.aws.amazon.com/xray/latest/devguide/index.html

User Guide: https://docs.aws.amazon.com/xray/latest/devguide/index.html

API Reference: https://docs.aws.amazon.com/xray/latest/api/

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: https://aws.amazon.com/xray/pricing/

Cost Optimization

AWS X-Ray offers several cost optimization features including customizable sampling rules that control trace volume without code changes, defaulting to 1 request per second plus 5% of additional requests. The generous free tier provides 100,000 traces recorded and 1,000,000 traces retrieved monthly at no cost. Customers can optimize costs by adjusting sampling rates based on application criticality, using annotations strategically to reduce indexed data, and leveraging the built-in sampling algorithm that automatically reduces trace volume while maintaining statistical significance for analysis.

Savings Considerations

Primary cost savings opportunities include leveraging the substantial free tier that covers many small to medium applications entirely, implementing intelligent sampling strategies that capture representative traces without recording every request, and using X-Ray's pay-as-you-go model to avoid upfront commitments. Costs scale predictably with application traffic, and the service eliminates the need for separate tracing infrastructure management. Compared to building custom distributed tracing solutions, X-Ray provides significant operational cost savings through reduced development, maintenance, and infrastructure overhead while delivering enterprise-grade capabilities.

Service Name

Kiro AI IDE

Service Overview

Kiro is AWS's native AI-powered development environment and integrated development environment (IDE) that enables developers to build applications using natural language specifications and AI assistance. It provides a comprehensive development experience with integrated access to AWS services, documentation, and real-time AI assistance. Kiro excels at contextual code analysis, infrastructure code generation, visual project management, and collaborative development workflows. The platform supports both CLI and IDE interfaces, leveraging the Model Context Protocol (MCP) to connect with specialized servers and extend capabilities with domain-specific knowledge and functionality.

Key Use Cases

- Cloud modernization and application migration: Automating assessment, planning, and implementation phases for migrating legacy applications to AWS
- Infrastructure optimization: Finding cost optimization opportunities and automatically applying recommendations to Infrastructure as Code
- AI agent development: Building and deploying production-ready AI agents with integrated Amazon Bedrock AgentCore support
- Database development: Assisting SQL Server professionals with query optimization, schema analysis, and performance tuning
- Architecture diagram generation: Creating AWS architecture diagrams using natural language prompts and MCP servers
- Citizen development: Enabling business users to build applications using natural language without traditional coding skills

Features

- **AI-Powered Development**: Natural language code generation and contextual analysis with real-time AI assistance
- **Model Context Protocol Integration**: Connects with specialized MCP servers for extended domain-specific capabilities
- **Infrastructure Code Generation**: Automated generation of Infrastructure as Code with AWS best practices
- **Visual Project Management**: Comprehensive project visualization and management capabilities
- **Cost Optimization**: Automated identification and application of cost optimization recommendations

- **Collaborative Workflows**: Multi-user collaboration features for development teams
- **Spec-Driven Development**: Requirements gathering, design, implementation planning, and execution workflow

Value Proposition

Kiro transforms traditional development workflows by enabling natural language programming and automated code generation, significantly reducing development time from weeks to hours. The platform integrates seamlessly with AWS services and follows AWS best practices automatically, ensuring secure and optimized deployments:

- Key advantages include citizen development capabilities that allow business users to build applications without coding expertise, automated cloud modernization that reduces assessment time from weeks to days, and comprehensive AI assistance that handles complex infrastructure patterns
- Kiro's spec-driven development methodology provides structured requirements, comprehensive design, systematic implementation, and built-in quality through testing and validation, making it ideal for organizations seeking rapid application development and modernization

Service Integration

Kiro integrates extensively with core AWS services through its MCP architecture and built-in connectors. Primary integrations include Amazon Bedrock AgentCore for AI agent runtime and deployment, AWS cost management tools for optimization recommendations, Amazon CloudWatch for monitoring and observability, Amazon RDS and Aurora for database development assistance, and AWS Lambda for serverless application development:

- The platform also connects with AWS documentation services, AWS Pricing API for cost analysis, Amazon S3 for storage management, and AWS Identity Center for authentication
- Kiro works with Infrastructure as Code tools like AWS CDK and CloudFormation, and supports integration with AWS Transform for application modernization and migration workflows

Onboarding and Offboarding

Customers can get started with Kiro by downloading the Kiro CLI or accessing the Kiro IDE through AWS. Initial setup requires configuring AWS credentials and setting up MCP servers for extended functionality:

- The platform offers multiple entry points including direct download, integration with existing IDEs, and web-based access

- Getting started involves installing required dependencies (Python 3.10, specific tools based on use case), configuring MCP servers in the settings file, and connecting to AWS services
- Kiro provides guided tutorials and documentation for common workflows including cloud modernization, cost optimization, and AI agent development
- The initial setup typically takes just a few hours, and users can begin with pilot projects to familiarize themselves with the agentic development approach

Getting Started Guide: Based on available documentation, specific getting started guide URLs are not explicitly provided in the search results, but users are directed to AWS documentation and blog posts for setup instructions

Data Removal

Offboarding information for Kiro is not explicitly detailed in the available documentation. As an AI development environment, data removal would likely involve clearing project workspaces, disconnecting MCP server integrations, removing AWS credential configurations, and deleting any locally stored project artifacts. Users should follow standard practices for removing development environment data and configurations.

Backup/Restore and Disaster Recovery

Backup and disaster recovery capabilities for Kiro are not explicitly documented in the available information. As a development environment, backup considerations would primarily focus on project workspaces, configuration files, and integration settings. Users would rely on standard development practices for code versioning and project backup.

Backup Capabilities

Specific backup capabilities for Kiro are not detailed in the available documentation. The platform likely relies on standard development environment backup practices and integration with version control systems for project preservation.

Disaster Recovery

Disaster recovery support for Kiro is not explicitly documented. As a development environment, disaster recovery would involve standard practices for development tool restoration and reconfiguration of MCP integrations and AWS connections.

Recovery Objectives

Specific RPO and RTO objectives for Kiro are not documented in the available information. Recovery objectives would depend on the underlying infrastructure hosting the development environment and standard development workflow recovery practices.

Service Constraints

Service Quotas: Service quotas documentation specific to Kiro is not available in the search results

FAQ: FAQ documentation specific to Kiro is not available in the search results

Technical Requirements

Service Documentation: Based on search results, Kiro documentation is primarily available through AWS blog posts and integration guides rather than centralized service documentation

User Guide: Specific user guide links are not provided in the search results, with information distributed across various AWS blog posts and tutorials

API Reference: API reference documentation specific to Kiro is not available in the search results

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: Specific pricing information for Kiro is not explicitly detailed in the search results, though costs may be associated with AI model usage and MCP server operations

Cost Optimization

Kiro offers unique cost optimization features including automated identification of optimization opportunities through MCP integration with AWS cost management tools, automatic application of recommendations to Infrastructure as Code, natural language creation of cost-optimized Service Control Policies, and embedded cost optimization rules in development workflows. The platform can analyze entire infrastructure files and suggest comprehensive optimizations including storage tier improvements, networking enhancements, compute right-sizing, auto-scaling configurations, and CloudWatch Logs retention policies. Kiro's context system allows organizations to embed their cost optimization best practices directly into the development workflow, ensuring every new resource is optimized from deployment.

Savings Considerations

Main cost savings come from dramatically reduced development time, with projects moving from weeks to hours through AI-assisted development and automated code generation. Kiro eliminates manual assessment and planning phases in modernization projects, reducing assessment time from weeks to days:

- The platform's automated optimization recommendations can identify significant cost savings opportunities across EC2 instances, auto-scaling groups, storage configurations, and other AWS resources
- Additional savings include reduced need for specialized development resources through citizen development capabilities, decreased maintenance overhead through automated best practices implementation, and prevention of cost overruns through embedded optimization rules and governance frameworks that are built into the development process from the start