

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