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