Amazon Glue

A serverless data integration service to discover, prepare, and integrate all your data at any scale

Discover, prepare, move, and integrate data from multiple sources with AWS Glue. AWS Glue is a fully managed extract, transform, and load (ETL) service.
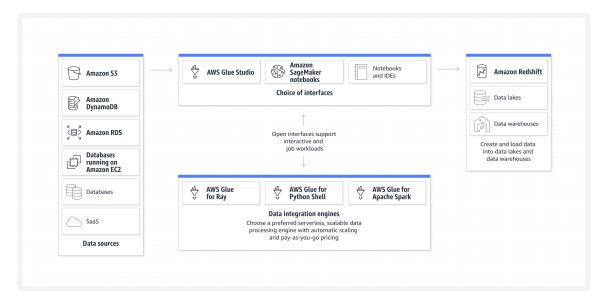
AWS Glue can help you discover and catalog metadata from various data sources for analytics and data processing. With AWS Glue, you can build data lakes, modernize data warehouses, or use machine learning and analytics at scale.

Concepts

- **Crawlers:** Detect new data, extract metadata, and populate the *Data Catalog*.
- **AWS Glue Data Catalog:** The persistent metadata store in AWS Glue.
- **Job:** The business logic that is required to perform ETL (extract, transform, and load) work.
- **Trigger:** Initiates an ETL job.
- **Script:** Code that defines an ETL job to extract data from sources, transform it, and load it into targets.

How it Works

AWS Glue uses other services to orchestrate your ETL jobs to build data warehouses and data lakes and generate output streams. AWS Glue calls API operations to transform your data, create runtime logs, store your job logic, and create notifications to help you monitor your job runs.

Use Cases

Data Discovery
Search across all AWS datasets or connect to your data source.

Streaming
Derive insights from data that comes in real-time.

Log Processing
Monitor issues, common patterns, trends, and more.

Data Cataloging
Create a centralized repository of metadata.

Data Transformation
Clean and transform streaming data in transit.