AWS Glue

is a fully-managed, pay-as-you-go, extract, transform, and load (ETL) service that automates the time-consuming steps of data preparation for analytics.

AWS Glue automatically discovers and profiles data via the Glue Data Catalog, recommends and generates ETL code to transform your source data into target schemas.

AWS Glue runs the ETL jobs on a fully managed, scale-out Apache Spark environment to load your data into its destination.

AWS Glue also allows you to setup, orchestrate, and monitor complex data flows.

You can create and run an ETL job with a few clicks in the AWS Management Console.

Simply point AWS Glue to your data stored on AWS, and AWS Glue discovers data and stores the associated metadata (e.g. table definition and schema) in the AWS Glue Data Catalog.

Once cataloged, data is immediately searchable, queryable, and available for ETL.

AWS Glue consists of a Data Catalog which is a central metadata repository, an ETL engine that can automatically generate Scala or Python code, and a flexible scheduler that handles dependency resolution, job monitoring, and retries.

Together, these automate much of the undifferentiated heavy lifting involved with discovering, categorizing, cleaning, enriching, and moving data, so you can spend more time analyzing your data.

**AWS Glue Crawlers**

You can use a crawler to populate the AWS Glue Data Catalog with tables.

This is the primary method used by most AWS Glue users.

A crawler can crawl multiple data stores in a single run.

Upon completion, the crawler creates or updates one or more tables in your Data Catalog. Extract, transform, and load (ETL) jobs that you define in AWS Glue use these Data Catalog tables as sources and targets.

The ETL job reads from and writes to the data stores that are specified in the source and target Data Catalog tables.

AWS Glue crawlers connect to a source or target data store, progress through a prioritized list of classifiers to determine the schema for the data, and then creates metadata in the AWS Glue Data Catalog.

The metadata is stored in tables in a data catalog and used in the authoring process of ETL jobs.

You can run crawlers on a schedule, on-demand, or trigger them based on an event to ensure that your metadata is up-to-date.

AWS Glue automatically generates the code to extract, transform, and load data.

Simply point AWS Glue to a source and target, and AWS Glue creates ETL scripts to transform, flatten, and enrich the data.

The code is generated in Scala or Python and written for Apache Spark.

AWS Glue helps clean and prepare data for analysis by providing a Machine Learning Transform called FindMatches for deduplication and finding matching records.

**Use Cases**

Use AWS Glue to discover properties of data, transform it, and prepare it for analytics.

Glue can automatically discover both structured and semi-structured data stored in data lakes on [**Amazon S3**](https://aws.amazon.com/s3/), data warehouses in [**Amazon Redshift**](https://aws.amazon.com/redshift/), and various databases running on AWS.

It provides a unified view of data via the Glue Data Catalog that is available for ETL, querying and reporting using services like [**Amazon Athena**](https://aws.amazon.com/athena/), [**Amazon EMR**](https://aws.amazon.com/emr/), and [**Amazon Redshift Spectrum**](https://aws.amazon.com/redshift/).

Glue automatically generates Scala or Python code for ETL jobs that you can further customize using tools you are already familiar with.

AWS Glue is serverless, so there are no compute resources to configure and manage.

<https://digitalcloud.training/aws-analytics-services/>

<https://digitalcloud.training/aws-migration-services/>

<https://digitalcloud.training/amazon-athena/>