AWS Glue

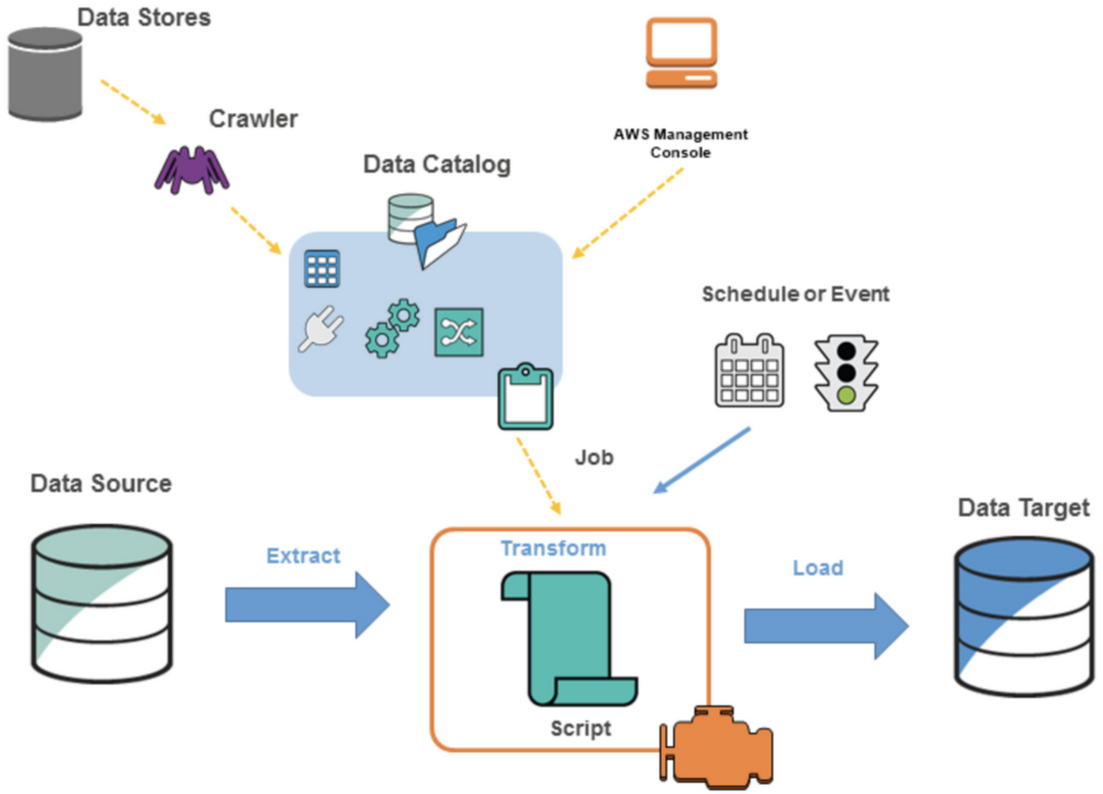
What is AWS Glue?

Fully Managed ETL Service

A Spark ETL Engine

Consists of a Central Metadata Repository - Glue Data Catalog

Flexible Scheduler



Why use AWS Glue?

AWS Glue offers a fully managed serverless ETL Tool. This removes the overhead, and barriers to entry, when there is a requirement for a ETL service in AWS.

AWS Glue - Setup Work

AWS Glue Data Catalog

Glue Data Catalog

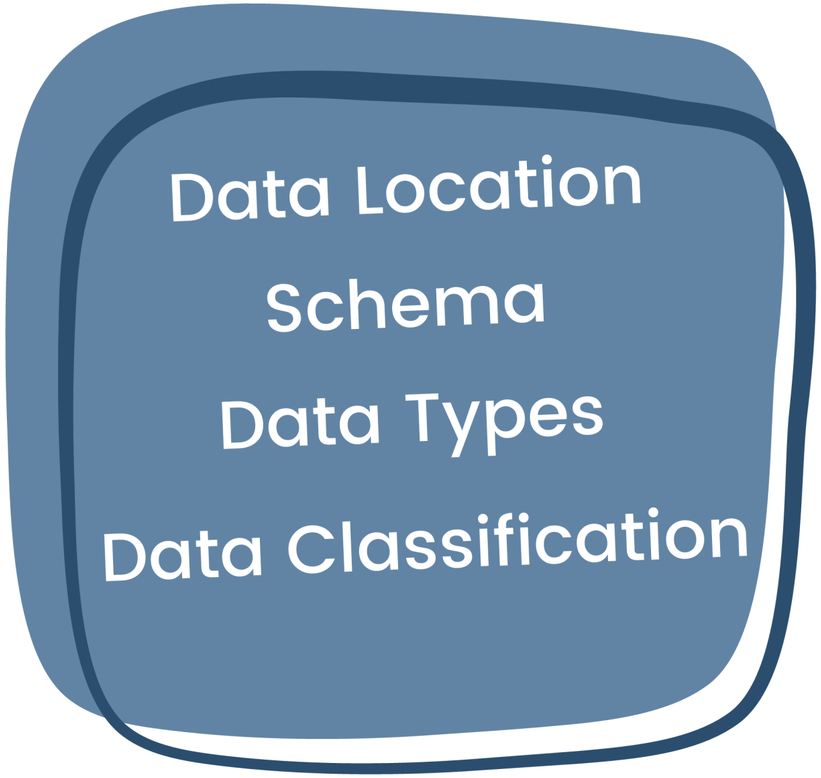
Persistent Metadata Store

It is a managed service that lets you store, annotate, and share metadata which can be used to query and transform data

One AWS Glue Data Catalog per AWS region

Identity and Access Management (IAM) policies control access

Can be used for data governance

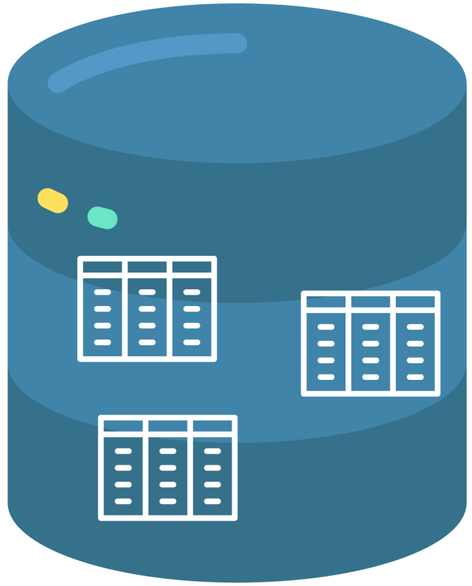


Examples of Meta Data

AWS Glue Databases

AWS Glue Databases

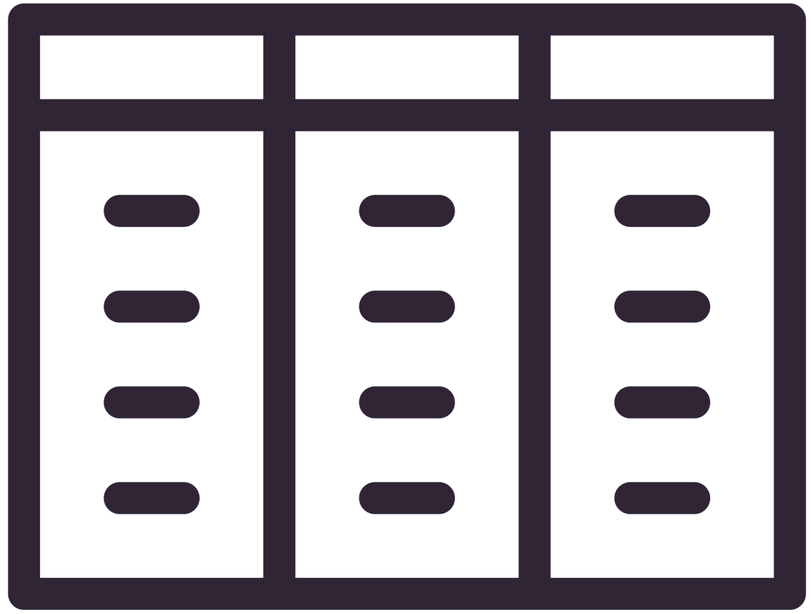
A set of associated Data Catalog table definitions organized into a logical group.



AWS Glue Tables

AWS Glue Tables

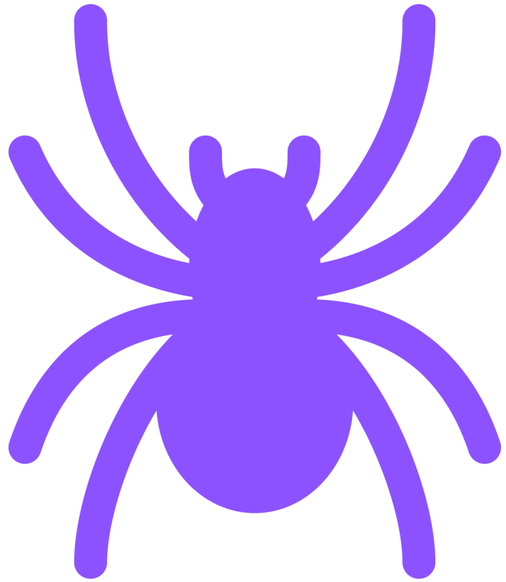
The metadata definition that represents your data. The data resides in its original store. This is just a representation of the schema.



AWS GLUE CRAWLERS

AWS Glue Crawler

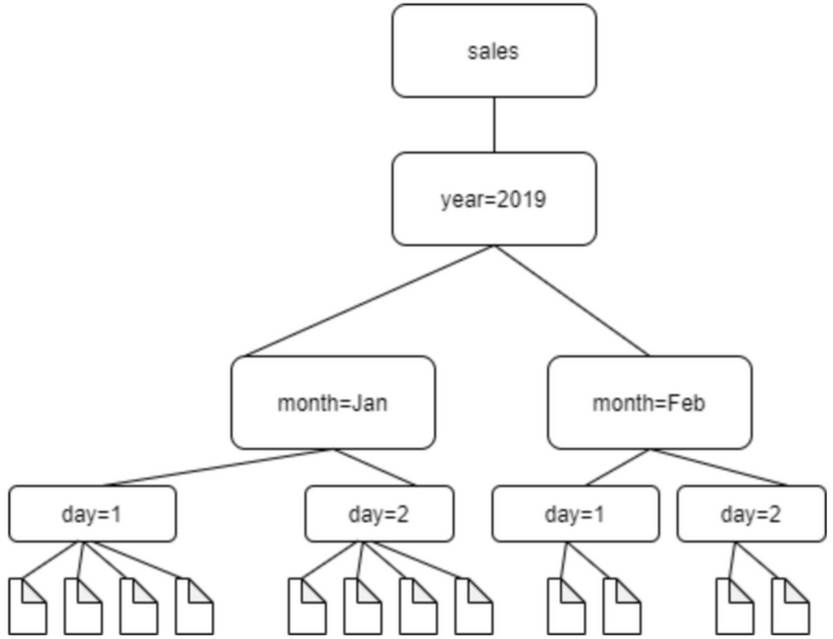
A program that connects to a data store (source or target), progresses through a prioritized list of classifiers to determine the schema for your data, and then creates metadata tables in the AWS Glue Data Catalog.



PARTITIONS IN AWS

PARTITIONS

Folders where data is stored on S3, which are physical entities, are mapped to partitions, which are logical entities i.e. Columns in the Glue table.

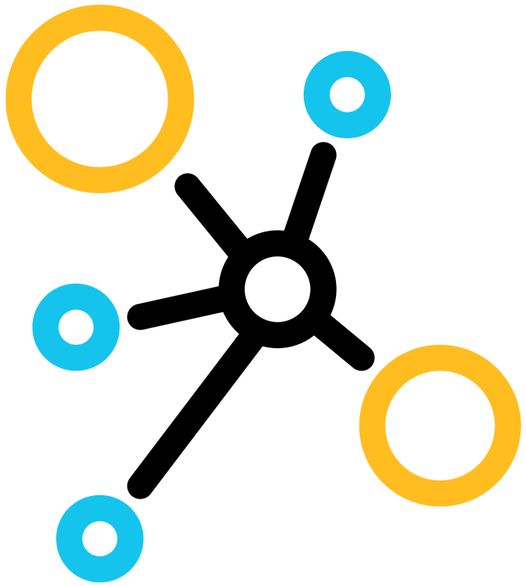


S3://sales/year=2019/month=Jan/day=1
S3://sales/year=2019/month=Jan/day=2
S3://sales/year=2019/month=Feb/day=1
S3://sales/year=2019/month=Feb/day=2

AWS Glue Connections

AWS Glue Connections

A Data Catalog object that contains the properties that are required to connect to a particular data store.



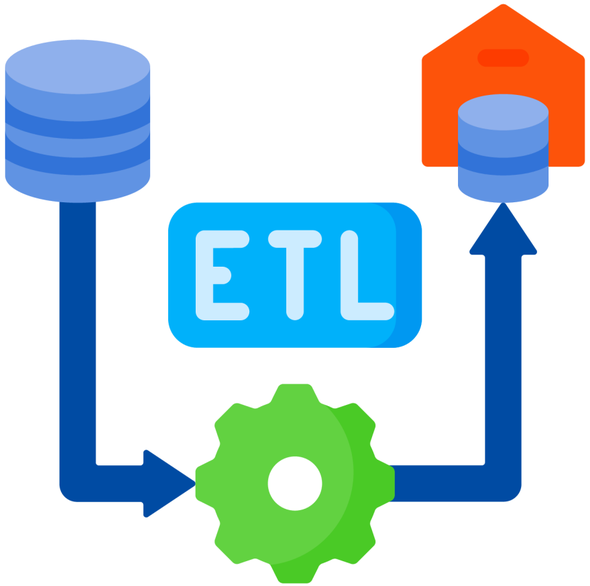
AWS Glue ETL

AWS Glue ETL

AWS Glue ETL supports extracting data from various sources, transforming it to meet your business needs, and loading it into a destination of your choice.

AWS Glue ETL Engine

Apache Spark engine to distribute big data workloads across worker nodes



AWS Glue DPUs

1 DPU is equivalent to 4 vCPUs and 16 GB memory.

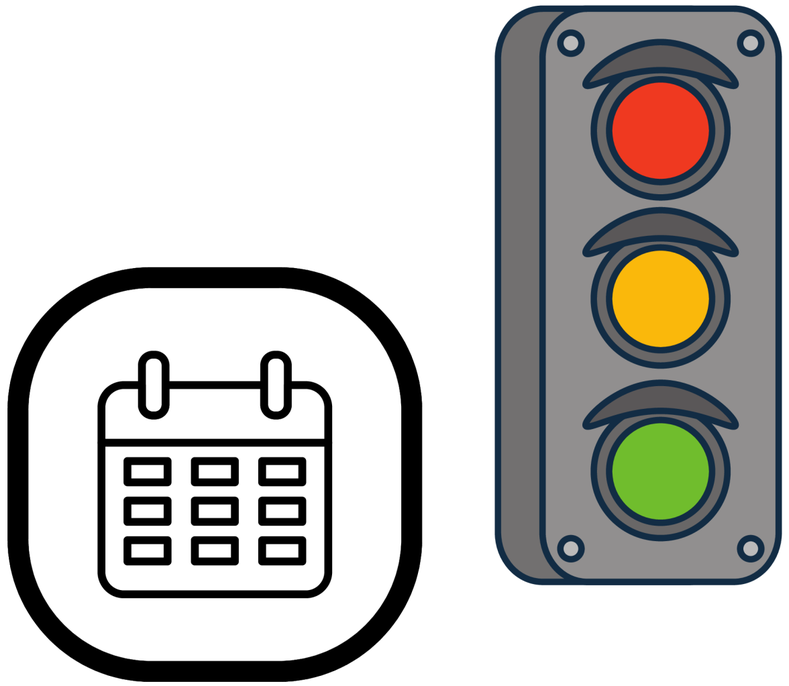
AWS Glue Bookmarks

Tracks data that has already been processed during a previous run of an ETL job by persisting state information from the job run

AWS Glue Scheduling

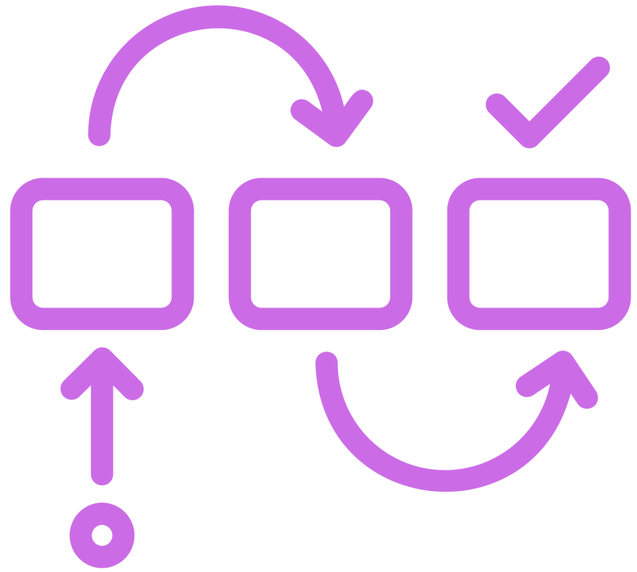
AWS Glue Triggers

Initiates an ETL job. Triggers can be defined based on a scheduled time or an event.



AWS Glue Workflow

Create and visualize complex extract, transform, and load (ETL) activities involving multiple crawlers, jobs, and triggers



AWS Glue Scheduling (Others)

Apache Airflow

AWS Step Functions

Amazon Event Bridge

AWS Glue Data Quality

AWS GLUE DATA Quality

Monitor the quality of your data by Data Quality Definition Language (DQDL) using DeeQu

AWS Glue Data Brew

AWS GLUE DATA Brew

Visual data preparation tool that makes it easier for data analysts and data scientists to clean and normalize data