AWS Glue

AWS Glue

- AWS Glue is a serverless data integration service.
- It provides all the capabilities needed for data integration, so you can start analyzing your data and putting it to use in minutes instead of months.

Catalog

Crawler

AWS Glue
Data Catalog

Process

AWS Glue Studio

AWS Glue
Development Endpoints

Glue Workflow

DevOps

Serverless Spark

Dynamic Frames

Glue Streaming

Quality

Machine Learning
powered data cleansing

Books
Catalog

CATALOG

Data Catalog

What Should a Data Catalog Offer?

Glue Catalog

- **Catalog** - An index to the location, schema, and runtime metrics of your data.
- **Databases** - A set of associated table definitions, organized into a logical group.
- **Tables** - The metadata definition that represents your data, including its schema.

The AWS Glue Data Catalog is an Apache Hive metastore-compatible catalog.

Reference:

FAQs

AWS Glue

What?

- AWS Glue is a serverless data integration service that makes it easy to discover, prepare, and combine data for analytics, machine learning, and application development.

Why?

- AWS Glue provides all the capabilities needed for data integration, so you can start analyzing your data and putting it to use in minutes instead of months.

When?

- You should use AWS Glue to discover properties of the data you own, transform it, and prepare it for analytics.
- When you need unified catalog to find data across multiple data stores, explore data with self-service visual data preparation or create, run, and monitor ETL jobs without coding.

Where?

- AWS Glue is a Regional service.

Who?

- AWS Glue provides a managed ETL service that runs on a serverless Apache Spark environment. This allows you to focus on your ETL job and not worry about configuring and managing the underlying compute resources.

How?

- You define jobs in AWS Glue to accomplish the work that's required to extract, transform, and load (ETL) data from a data source to a data target.
- AWS Glue can generate a script to transform your data. Or, you can provide the script in the AWS Glue console or API.

Category:

Analytics

How much?

- With AWS Glue, you pay an hourly rate, billed by the second, for crawlers and ETL jobs. For the AWS Glue Data Catalog, you pay a monthly fee for storing and accessing the metadata. For development endpoint, you pay an hourly rate, billed per second.
- For AWS Glue DataBrew, the interactive sessions are billed per session and the DataBrew jobs are billed per minute.

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.
- This service uses the Apache Spark engine to distribute workloads across worker nodes, enabling faster transformations with in-memory processing.

AWS Glue Studio

- AWS Glue Studio is a graphical interface that makes it easy to create, run, and monitor data integration jobs in AWS Glue.
- You can visually compose data transformation workflows and seamlessly run them on the Apache Spark–based serverless ETL engine in AWS Glue.
- You can also run your existing Scala or Python code on AWS Glue.
- AWS Glue supports ETL on streams from Amazon Kinesis Data Streams, Apache Kafka, and Amazon MSK.
- You can start the job creation process from a blank canvas, notebook or script editor.

AWS Workshop

- MySQL database – Data Model

We will join three tables and do an aggregation to identify which product type and in which zip code generated the highest sales.

- Workshop Architecture

AWS Workshop

- Workshop Architecture

- MySQL database – Data Model

We will join three tables and do an aggregation to identify which product type and in which zip code generated the highest sales.

Running an AWS Glue ETL Job

- Worker Type
  - Predefined capacity to run your job
  - The resources available on AWS Glue workers are measured in DPUs (Glue Data Processing Unit).
  - A DPU is a relative measure of processing power that consists of 4 vCPUs of compute capacity and 16 GB of memory.

- Glue Version
  - Glue 4.0
    - Spark 3.3.0, Python 3.10,
  - Glue 3.0
    - Spark 3.1.1, Python 3.7

AWS Glue Bookmark

Bookmark

AWS Glue – Job Bookmark

- AWS Glue tracks data that has already been processed during a previous run of an ETL job by persisting state information from the job run. This persisted state information is called a job bookmark.

- Job bookmarks help AWS Glue maintain state information and prevent the reprocessing of old data. With job bookmarks, you can process new data when rerunning on a scheduled interval.

- You can rewind your job bookmarks for your AWS Glue Spark ETL jobs to any previous job run. You can support data backfilling scenarios better by rewinding your job bookmarks to any previous job run, resulting in the subsequent job run reprocessing data only from the bookmarked job run.

AWS Glue Jobs Execution Types

AWS Glue Jobs – Execution Types

- AWS Glue jobs with various execution types
  - Standard
    - AWS Glue holds all the computing resources for the whole job run until it is completed.
  - Auto Scaling
    - AWS Glue will optimize costs and resource usage by dynamically scaling the number of workers up and down throughout the job run.
  - Flex Execution
    - AWS Glue jobs run on spare capacity in AWS instead of dedicated reserved or on-demand EC2 instances.

Job name	Type	Run status	Run time	Capacity	Worker type	DPU hours
NOAA Flex Execution	Glue ETL	Succeeded	13 minutes	10	G.1X	1.40 ($ 0.41)
NOAA Auto Scaling	Glue ETL	Succeeded	8 minutes	10	G.1X	1.12 ($ 0.49)
NOAA Basic	Glue ETL	Succeeded	8 minutes	10	G.1X	1.28 ($ 0.56)

Use Cases

- Standard execution:
  - Ideal for time-sensitive workloads that require fast job startup and dedicated resources.
  - Jobs that have downstream dependencies are better served by the standard execution class.
- Auto Scaling:
  - Batch jobs to process unpredictable amounts of data
  - Jobs containing driver-heavy workloads (for example, processing many small files)
  - Jobs containing multiple stages with uneven compute demands or due to data skews (for example, reading from a data store, repartitioning it to have more parallelism, and then processing further analytic workloads)
  - Jobs to write large amounts of data into data warehouses such as Amazon Redshift or read and write from databases.
- Flex execution:
  - Ideal for reducing the costs of time-insensitive workloads.
  - Nightly ETL jobs, or jobs that run over weekends for processing workloads
  - One-time bulk data ingestion jobs
  - Jobs running in test environments or pre-production workloads
  - Time-insensitive workloads where it’s acceptable to have variable start and end times

Glue Workflow

A complex ETL Process

AWS Glue – Workflows

- A workflow is a container for a set of related jobs, crawlers, and triggers.
- Using workflow, you can design a complex ETL activity that AWS Glue can execute and track as a single entity.
- Glue console provides a visual representation of a workflow as a graph.
- Key component of work flow
  - Trigger
    - Responsible to start the workflow (or crawler or job)
  - Node
    - Responsible to run job or a crawler

Glue Data Quality

Gourmet Meal requires quality Ingredients

Data Quality matters

AWS Glue Data Quality

Serverless

Rule recommendations

Data-at-rest and data-in-transit support

Powered by Deequ - an open-source library developed by Amazon to manage and monitor data quality

Few built-in rules

Rule types

Consistency

Column Correlation
Column Data Type
Column Exists
Column Length
Column Names Match Pattern

Accuracy

Column Values
Custom Rule with SQL support
Mean
Row Count
Standard Deviation
Sum
Distinct Value Count

Completeness

Completeness
Completeness with threshold
Entropy
Data Freshness

Integrity

Is Primary Key
Uniqueness
Uniqueness with threshold
Unique Value Ratio

Data-at-rest and data-in-transit support

Data Quality at Rest

Selects a dataset in the AWS Glue Data Catalog

AWS Glue Data Quality analyzes data and recommends rules

You refines rules to create finalized rules

AWS Glue Data Quality evaluates rules

Reviews results and alerts and takes appropriate action

Data Quality in transit

Selects an ETL job and adds AWS Glue Data Quality rules and actions

AWS Glue Data Quality evaluates rules

Review results and alerts and takes appropriate action

AWS Glue DataBrew

AWS Glue DataBrew

- AWS Glue DataBrew is a visual data preparation tool to prepare data with an interactive, spreadsheet-style UI and point-and-click visual interface without writing code.
- It is built for users who need to clean and normalize data for analytics and ML. Data analysts and data scientists are the primary users.

- DataBrew is available in a separate console view from AWS Glue. You can use it with or without using Glue Data Catalog.
- You can choose from over 250 prebuilt transformations to automate data preparation tasks.

Core concepts in DataBrew

- Project
  - The entire data preparation workspace in DataBrew
- Dataset
  - A collection of structured or semi-structured data
- Recipe
  - A set of data transformation steps; each step can contain many actions
- Job
  - A set of instructions to run a recipe or a data profile job
- Data lineage
  - The tracking of data in a visual interface to identify its origin
- Data profile
  - A summary view of the shape of your data

Immersion Day

Tutorial

Let’s brew Citibike data using AWS Glue DataBrew

CitiBike

19,000 bikes and over 1000 stations

AWS Glue
Additional Features

Store the credentials in AWS Secrets Manager for connections

AWS Glue > Connectors > Create connection

Step 1
Choose data source

Step 2
Configure connection

Step 3
Set properties

Step 4
Review and create

Configure connection

Connection details

Database instances
Provisioned Amazon Relational Database Service instances.

Choose one JDBC URL

Database name

Credential type

- ○ Username and password
- ◉ AWS Secrets Manager

AWS Secret - *optional* Info
Choose a secret from AWS Secrets Manager. AWS secrets eliminate hardcoding sensitive information.

Choose one secret

Use VPC Endpoints to access data in Amazon S3

- Make sure you configure Route Table entries
- Tighten security by using VPC endpoint policies

Glue Worker Type

- If jobs are taking too much time, try scaling the worker types (G.4X and G.8X)

AWS Glue Worker Type	DPU per Node	vCPU	Memory (GB)	Disk (GB)	Number of Spark Executors per Node	Number of Cores per Spark Executor
G.1X	1	4	16	64	1	4
G.2X	2	8	32	128	1	8
G.4X (new)	4	16	64	256	1	16
G.8X (new)	8	32	128	512	1	32

AWS Glue – Automatically identify partitions

- If needed create partitions using Create_Partition API

AWS Glue Schema Registry

- Allows you to centrally discover, control, and evolve data stream schemas.
- The Schema Registry allows disparate systems to share a schema for serialization and de-serialization.
- With AWS Glue Schema Registry, you can manage and enforce schemas on your data streaming applications using convenient integrations with Apache Kafka, Amazon Managed Streaming for Apache Kafka, Amazon Kinesis Data Streams, Amazon Managed Service for Apache Flink, and AWS Lambda.

```
{
  "type": "record",
  "namespace": "ABC_Organization",
  "name": "Employee",
  "fields": [
    {
      "name": "Name",
      "type": "string"
    },
    {
      "name": "Age",
      "type": "int"
    },
    {
      "name": "address",
      "type": {
        "type": "record",
        "name": "addressRecord",
        "fields": [
          {
            "name": "street",
            "type": "string"
```