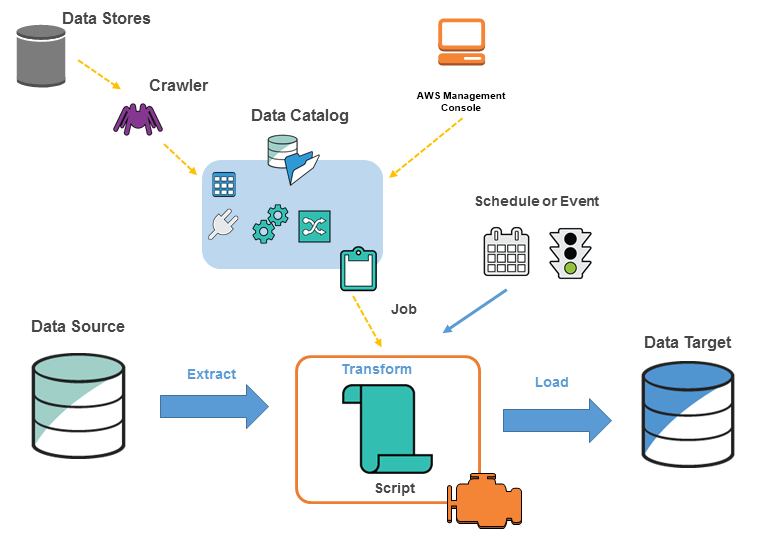
**Glue:**

1. **AWS Glue** is a fully managed extract, transform, and load (ETL) service provided by Amazon Web Services (AWS).
2. It helps users prepare and transform data for analytics, machine learning, and application development.
3. AWS Glue automates the data preparation process by discovering, cataloging, cleaning, enriching, and transforming data from various sources.



**Key Features of AWS Glue:**

1. **Data Integration:** Connects to a wide range of data sources including databases, data lakes, and SaaS applications.
2. **Data Catalog:** Maintains a centralized metadata repository to store and manage data schemas and definitions.
3. **ETL Automation:** Generates ETL scripts automatically using Apache Spark-based processing.
4. **Serverless:** No infrastructure management is needed as AWS Glue handles scaling and maintenance.
5. **Job Scheduling:** Supports job scheduling, monitoring, and retry mechanisms.
6. **Data Preparation Tools:** Includes AWS Glue Studio for visual ETL job creation and AWS Glue DataBrew for interactive data preparation.

**Fully Managed:**

* F**ully managed** means that AWS takes care of the operational tasks associated with running a service, so users don’t have to manage infrastructure, scaling, maintenance, or availability.
* This allows customers to focus on using the service and building applications instead of handling backend management tasks.

**Serverless Services:**

* Serverless services eliminate the need to manage servers entirely.
* AWS automatically provisions, scales, and shuts down resources as needed.
* Users only pay for the actual compute time or data processed.

**Use Case:**

**1) Run queries against S3 Data Lake:**

You can use AWS Glue to make your data available for analytics without moving your data.

**2) Analyse the log data in your data warehouse:**

Create ETL scripts to transform, flatten and enrich the data from source to target.

**3) Create event-driven ETL pipelines:**

As soon as new data becomes available in S3 you can run your ETL job by invoking Glue ETL jobs using an AWS Lambda Function.

**4) A unified view of your data across multiple data stores:**

With Glue Data Catalog you can quickly search and discover all your datasets and maintain the relevant metadata in one central repository.

**Glue Concepts:**

**1) AWS Glue Data Catalog:**

* A persistent metadata store.
* The data is used as sources & targets for your ETL Jobs.
* You can only use one data catalog per region.
* AWS Glue Data Catalog can be used as Hive metastore.
* It can contain database & tables.

**2) Database:**

* A set of associated table definitions organized into a logical group.
* A container for tables that define data from different data stores.
* If the database is deleted from the data catalog all the tables in the database gets deleted.
* A link to a local or shared database is called a database resource link.

**3) Data Store, Data Source & Data Target:**

* To persistently store your data in a repository you can use a data store like S3, RDS, Redshift, DynamoDB, JDBC etc.
* The **data source** is used as input to a process or transform.
* A location where the data store process or transform writes is called a **data target**.

**4) Table:**

* The metadata definition that represents your data.
* You can define tables using Json, Csv, Parquet, Avro & XML.
* You can use the table as the source or target in a job definition.
* A link to a local or shared table is called a table resource link.
* Tables can be added using crawler manually using Glue console or API CreateTable operation.
* A partitioned table describes an AWS Glue table definition of S3 folder.

**5) Connection:**

* It contains the properties that you need to connect the data.
* To store connection information for the data store you can add a connection using: **JDBC, RDS, Redshift, Document DB** etc.
* You can enable SSL connection for **JDBC, RDS, Redshift** and **Document DB**.

**What is SSL?**

* An SSL connection is a secure connection between a browser and a website that uses the Secure Sockets Layer (SSL) protocol to encrypt data.
* SSL prevents hackers from accessing or stealing information that is transferred between the two, such as financial or personal data.

**6) Crawler:**

* You can use crawlers to populate Glue Data Catalog with tables.
* Crawlers can crawl file-based and table-based data stores. It can crawl multiple tables in a single run.
* Crawlers determine the format, schema and associated properties of the raw data by classifying the data.
* Crawlers group the data into tables and partitions.
* It writes metadata to Glue Data Catalog.
* For incremental datasets you can use incremental crawls. It only crawls the folders that were added since the last crawler run.
* You can run a crawler on-demand or based on a schedule.
* Select the Logs link to view the results of the crawler. The link redirects you to CloudWatch Logs.

**7) Classifier:**

* It reads the data in data store.
* You can use a set of built-in classifiers or create a custom classifier. Custom classifier types are XML, JSON & CSV.
* By adding a classifier you can determine the schema of your table.

**8) AWS Glue Studio:**

* Visually author, run, view and edit your ETL Jobs.
* Diagnose, debug and check the status of your ETL Jobs.

**9) Job:**

* To perform ETL Works you need to create a job.
* When creating a job you need to provide data sources, targets and other transformations. The result will be generated in a PySpark Script and store the job information in the Glue Data Catalog.
* **Job Types:** Spark, Spark Streaming and Python Shell.
* **Job properties:**

*Job bookmarks* maintain the state information and prevent the reprocessing of old data.

*Job metrics* allows you to enable or disable the creation of CloudWatch metrics when the job runs.

*Security configuration* helps you define the encryption options of the ETL job.

*Worker type* is the predefined worker that is allocated when a job runs.

1. Standard
2. G.1X (memory-intensive jobs)
3. G.2X (jobs with ML transforms)

* *Max concurrency* is the maximum number of concurrent runs that are allowed for the created job. If the threshold is reached, an error will be returned.
* *Job timeout* (minutes) is the execution time limit.
* *Delay notification threshold* (minutes) is set if a job runs longer than the specified time. AWS Glue will send a delay notification via Amazon CloudWatch.
* *Number of retries* allows you to specify the number of times AWS Glue would automatically restart the job if it fails.
* *Job parameters* and *Non-overrideable Job parameters* are a set of key-value pairs.

**10) Script:**

* A script allows you to extract the data from sources, transform it, and load the data into the targets.
* You can generate ETL scripts using Scala or PySpark.
* AWS Glue has a script editor that displays both the script and diagram to help you visualize the flow of your data.

**11) Trigger:**

* It allows you to start manually/automatically one or more crawlers/ETL jobs.
* You can define triggers based on schedule, job events, and on-demand.
* You can also use triggers to pass job parameters. If a trigger starts multiple jobs, the parameters are passed on each job.

**15) Workflows:**

* It helps you orchestrate ETL jobs, triggers, and crawlers.
* Workflows can be created using the AWS Management Console or AWS Glue API.
* You can visualize the components and the flow of work with a graph using the AWS Management Console.
* Jobs and crawlers can fire an event trigger within a workflow.
* By defining the default workflow run properties, you can share and manage state throughout a workflow run.

**16) Pricing:**

* You are charged at an hourly rate based on the number of DPUs used to run your ETL job/ crawler.
* Data Catalog storage and requests:

1. You will be charged per month if you store more than a million objects.
2. You will be charged per month if you exceed a million requests in a month.