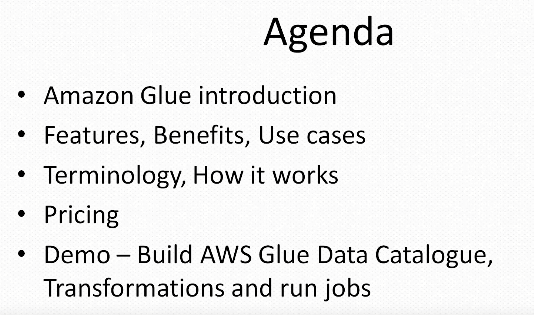
**AWS Glue-Managed ETL Service**

AWS Glue is a fully managed ETL (extract, transform, and load) service that makes it simple and cost-effective to categorize your data, clean it, enrich it, and move it reliably between various data stores.



**AWS Glue consists of**

1. a central metadata repository known as the **AWS Glue Data Catalog**
2. an **ETL engine** that automatically generates Python or Scala code
3. and a **flexible scheduler** that handles dependency resolution, job monitoring, and retries.

**Note:** AWS Glue is **serverless**, so there’s no infrastructure to set up or manage.

**How it Works:**

1. Select a data source and data target.
2. AWS Glue will generate ETL code in Scala or Python to extract data from the source, transform the data to match the target schema, and load it into the target.
3. You can edit, debug and test this code via the Console, in your IDE, or any notebook.

**Step1:** Built your Data Catalog

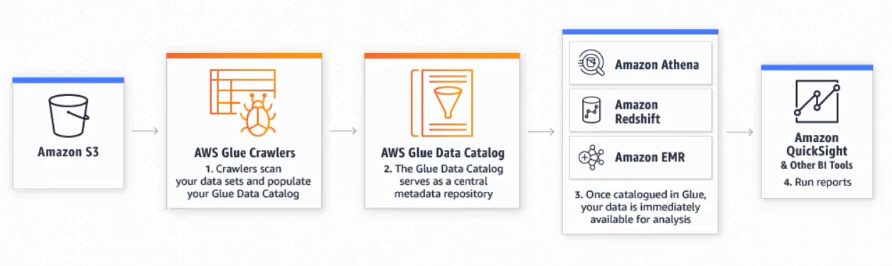
**Step2:** Generate and Edit Transformations

**Step3:** Schedule and Run Your Jobs

**Use Case: 1**

**Queries against an Amazon S3 Data Lake**

1. Data lake are an increasingly popular way to store and analyze both structured and unstructured data.
2. If you use an Amazon S3 data lake, AWS Glue can make all your data immediately available for analytics without moving the data.



**Notes:**

1. **AWS Glue Crawlers:** Crawlers scan your data sets and populate your Glue Data Catalog.
2. **AWS Glue Data Catalog:** The Glue Data Catalog serves as a Central metadata repository
3. **Amazon Athena, Amazon Redshift, Amazon EMR:** Once catalogued in Glue, your data is immediately available for analysis.
4. **Amazon QuickSight & Other BI Tools:** Run reports.

**Use Case: 2**

**Analyze Log Data in Your Data Warehouse**

1. Prepare your clickstream or process log data for analytics by cleaning, normalizing, and enriching your data sets using AWS Glue.
2. AWS Glue generates the schema for your semi-structured data, creates ETL code to transform, flatten, and enrich your data, and loads your data warehouse on a recurring basis.



**Notes:**

**Amazon RDS:** Enrich the data by joining with a relational database.

**Amazon Redshift:** Loads the data as tables to your data warehouse.

**Amazon QuickSight:**  To Run Reports.

**Amazon CloudWatch:** Logs and notifications are pushed to CloudWatch.

**AWS Glue ETL:** Schedule jobs to cleanse and flatten the data set.

**Use Case 3:**

**Unified View of Your Data Across Multiple Data Stores**

1. You can use the AWS Glue Data Catalog to quickly discover and search across multiple AWS data sets without moving the data.
2. Once the data is cataloged, it is immediately available for search and query using Amazon Athena, Amazon EMR, and Amazon Redshift Spectrum.



**Note:**

**AWS Glue ETL:** Read and Write metadata to the Glue Data Catalog.

**Amazon Athena, Redshift, EMR**: Access the Glue Data Catalog for ETL and analysis.

**Amazon QuickSight:** Run reports.

**AWS Glue Data Catalog:**  The Glue Data Catalog servers as a central metadata repository.

**Use Case 3:**

**Event-driven ETL Pipelines**

1. AWS Glue can run your ETL jobs based on an event, such as getting a new data set.
2. For example, you can use an AWS Lambda function to trigger your ETL jobs to run as soon as new data becomes available in Amazon S3.
3. You can also register this new dataset in the AWS Glue Data Catalog as part of your ETL jobs.



**Amazon S3:** New data is loaded into S3.

**AWS Lambda:** Trigger one or more Glue ETL jobs.

**AWS Glue ETL:** Schedule ETL jobs to transform and load the data.

**Amazon Redshift, Amazon S3:** Glue loads transformed data into the target data store.

**Amazon CloudWatch:** Logs and notifications are pushed to CloudWatch.

**AWS Glue Data Catalog:** The Glue Data Catalog servers as a central metadata repository.

