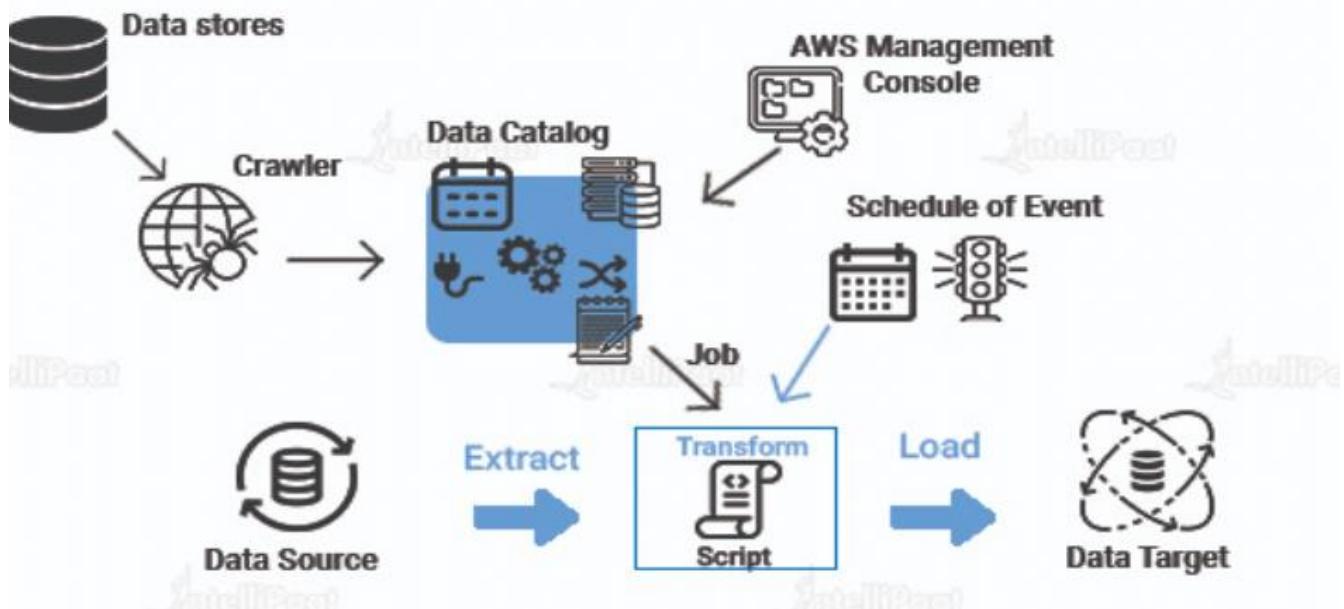


## AWS GLUE

AWS Glue is a fully managed serverless data integration service provided by Amazon Web Services. It's designed to make it easy to prepare, transform, and load (ETL) data for analytics, machine learning, and application development. Here's a quick breakdown of its key components and use cases:



### **Key Features of AWS Glue**

1. Serverless: No infrastructure to manage. AWS provisions and scales compute resources automatically.
2. ETL Jobs: You can create jobs that extract, transform, and load data using Python or Scala (typically Spark-based).
3. Glue Studio: A visual interface for building and running ETL jobs with drag-and-drop functionality.
4. Glue Data Catalog: A central metadata repository to store schema information about your datasets.
5. Job Triggers & Workflows: Automate ETL pipelines with triggers, job dependencies, and conditional logic.
6. Crawlers: Automatically scan data sources to detect schema and update the Data Catalog.
7. Data Brew: A no-code visual tool for data preparation for business analysts and data scientists.

## Common Use Cases

- Data warehouse ETL (e.g., from S3 to Redshift or Snowflake).
- Data lake management and transformation.
- Metadata cataloguing across datasets in S3, RDS, etc.
- Real-time data processing with AWS Glue Streaming (Kafka, Kinesis).
- Batch transformations on large datasets with Spark.

## Step-by-Step Guide:

### 1. Log in to AWS Management Console:

Go to AWS Glue service.

The screenshot shows the AWS Glue Welcome page. On the left, there's a sidebar with navigation links for AWS Glue, including 'Getting started', 'ETL jobs', 'Data Catalog', 'Data Integration and ETL', and 'Legacy pages'. Below that are 'What's New' and 'Documentation'. The main content area has a dark blue header with the text 'Welcome to AWS Glue' and a sub-header 'Get started by setting up your account and users, cataloging your data, and building ETL jobs to prepare data for analytics.' Below the header are three cards: 'Prepare your account for AWS Glue' (with a shield icon), 'Catalog and search for datasets' (with a magnifying glass icon), and 'Move and transform data' (with a circular arrow icon). Each card has a call-to-action button at the bottom.

### 2. Create a New Database:

Go to Data Catalog → Databases → Add Database.

Name: **retail\_sales\_db**

Description: "Retail sales data for analysis"

Go to AWS glue, open data catalog → Tables → create a table

The screenshot shows the AWS Data Catalog Tables page. On the left, there's a sidebar with navigation links for 'Data Catalog tables', 'Data connections', 'Workflows (orchestration)', 'Zero-ETL integrations', 'Data Catalog', 'Databases', 'Tables', 'Stream schema registries', 'Schemas', 'Connections', 'Crawlers', 'Classifiers', 'Catalog settings', and 'Data Integration and ETL'. The main content area has a banner at the top announcing new optimization features for Apache Iceberg tables. Below the banner, there's a section titled 'Tables' with a sub-section 'Tables (1)'. It shows a table with one row: 'table1' in the Name column, 'retail\_data' in the Database column, 's3://mynytaxid' in the Location column, and 'Parquet' in the File format column. There are buttons for 'Delete', 'Add tables using crawler', and 'Add table'.

### 3. Set Up IAM Role:

Ensure a role exists with these policies:

**AWS Glue Service Role**

**AmazonS3ReadOnlyAccess**

If not, create a new role in the IAM console.

The screenshot shows two screenshots of the AWS IAM service. The top screenshot displays the 'readaccess' role details, including its ARN (arn:aws:iam::239710306715:role/readaccess), last activity (April 29, 2025, 13:08 UTC-04:00), and maximum session duration (1 hour). The 'Permissions' tab is selected, showing one attached policy: 'AmazonS3ReadOnlyAccess'. The bottom screenshot shows the confirmation message 'Role readaccess created.' after a new role was created. It also shows the 'Roles Anywhere' section, which provides instructions for authenticating non-AWS workloads using X.509 certificates or temporary credentials.

### 4. Create a Crawler:

Go to Crawlers → Add Crawler.

Name: sales\_data\_crawler

Data Store: S3

Path: s3://retail-data-bucket/sales\_data/

IAM Role: Use the role created earlier.

Output Database: Select retail\_sales\_db

Run Frequency: Run on demand (for now)

## Review and create the crawler.

Open AWS glue → data catalog → crawlers →create crawler

The screenshot shows the 'Set crawler properties' step in the 'Add crawler' wizard. On the left, a sidebar lists various AWS services like ETL jobs, Data Catalog tables, and Data connections. The main panel has a title 'Set crawler properties' and a sub-section 'Crawler details'. It includes fields for 'Name' (set to 'sales\_data\_crawler') and 'Description - optional' (with placeholder 'Enter a description'). A section for 'Tags - optional' is also present. The sidebar on the left shows the progress: Step 1 (Set crawler properties) is completed (blue circle), while Step 2 through Step 5 are still pending (light blue circles).

The screenshot shows the 'Add data source' step in the 'Add crawler' wizard. The sidebar shows the progress: Step 1 is completed, Step 2 is selected (blue circle), and Steps 3 through 5 are pending. The main panel shows a 'Data source' dropdown set to 'S3'. Below it is a 'Network connection - optional' section with a 'Clear selection' and 'Add new connection' button. Under 'Location of S3 data', the 'In this account' option is selected. The 'S3 path' field contains 's3://mynyctaxidata'. The 'Subsequent crawler runs' section has a 'Cancel' button and an 'Add an S3 data source' button. A preview window on the right shows a list of existing data sources.

The screenshot shows the 'Configure security settings' step in the 'Add crawler' wizard. The sidebar shows the progress: Step 1 is completed, Step 2 is selected (blue circle), and Steps 3 through 5 are pending. The main panel has a 'IAM role' section with an 'Existing IAM role' dropdown set to 'readaccess'. It includes 'Create new IAM role' and 'Update chosen IAM role' buttons. Below is a 'Lake Formation configuration - optional' section with a checkbox for 'Use Lake Formation credentials for crawling S3 data source'. The sidebar on the left shows the progress: Step 1 is completed, Step 2 is selected, and Steps 3 through 5 are pending.

The screenshot shows the AWS Glue 'Add crawler' wizard. The left sidebar shows navigation links like 'Getting started', 'ETL jobs', 'Visual ETL', etc., and a 'Data Catalog' section with 'Crawlers' selected. The main area has a vertical progress bar with steps: Step 2 (Choose data sources and classifiers), Step 3 (Configure security settings), Step 4 (Set output and scheduling, which is highlighted with a blue circle), and Step 5 (Review and create). The 'Output configuration' section includes a 'Target database' dropdown set to 'retail\_data', a 'Table name prefix - optional' input field, and a 'Maximum table threshold - optional' input field with the placeholder 'Type a number greater than 0'. There is also a 'Crawler schedule' section.

The screenshot shows the 'sales\_data\_crawler' details page. The left sidebar is identical to the previous screenshot. The main area displays a green success message: 'One crawler successfully created. The following crawler is now created: "sales\_data\_crawler"'. Below this, the crawler properties are listed in a table:

Name	IAM role	Database	State
sales_data_crawler	readaccess	retail_data	READY

Other properties shown include 'Description' (empty), 'Security configuration' (empty), 'Lake Formation configuration' (empty), and 'Table prefix' (empty). There is also a 'Maximum table threshold' section with a value of '-'. A 'Crawler runs' tab is visible at the bottom.

## 5. Run the Crawler:

Select the crawler and click Run Crawler.

Wait for it to complete and check the Tables under `retail_sales_db`.

The screenshot shows the AWS Glue interface with the 'Crawlers' section selected. A single crawler named 'sales\_data\_c...' is listed, showing it is in a 'Ready' state with a success rate of 100% and last run on April 29, 2025. The interface includes a search bar, sorting options, and buttons for action, run, and creating a new crawler.

## 6. Verify the Metadata:

Go to the Tables section in the Data Catalog.

Review the schema — columns, data types, and partition

The screenshots show the AWS Glue Data Catalog Tables interface. The top screenshot displays the 'Schema' tab for a table named 'table1'. It shows the table's details: Name (table1), Classification (Parquet), Location (s3://mynyctaxidata/), and Connection (-). The bottom screenshot displays the 'Partitions' tab for the same table. It shows the table's details again and then lists a single partition: Database (retail\_data), Location (s3://mynyctaxidata/), and Connection (-).