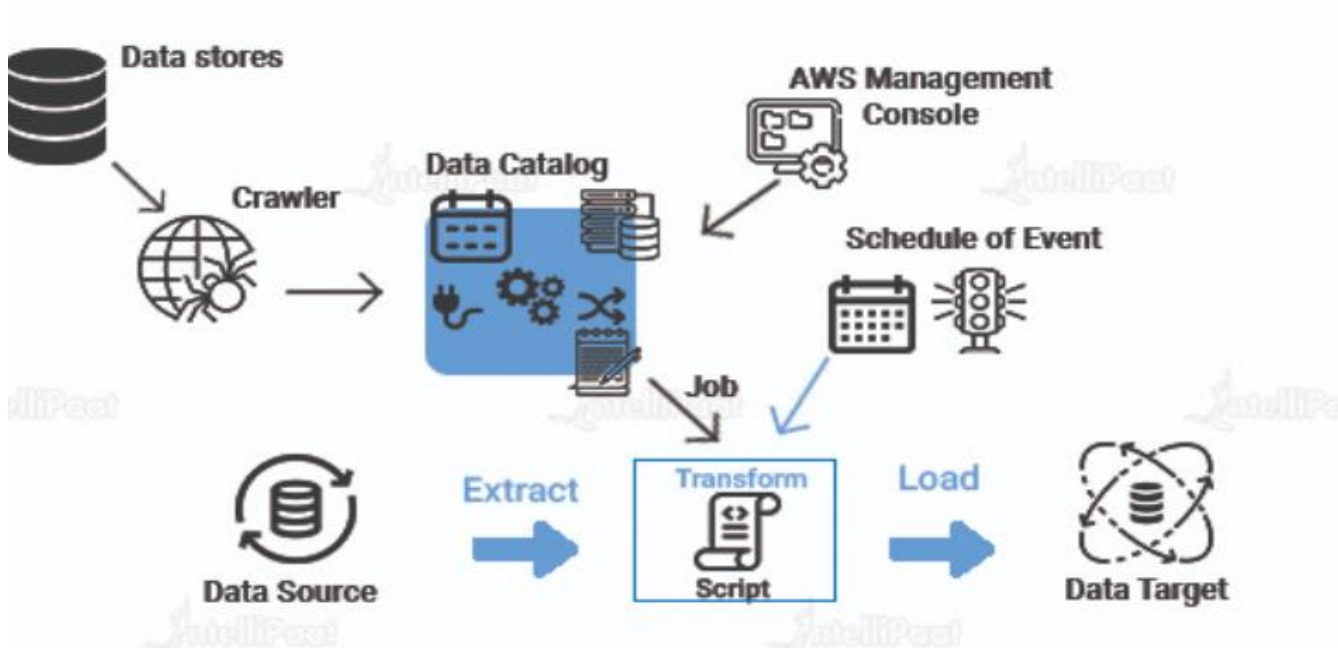


AWS GLUE

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.

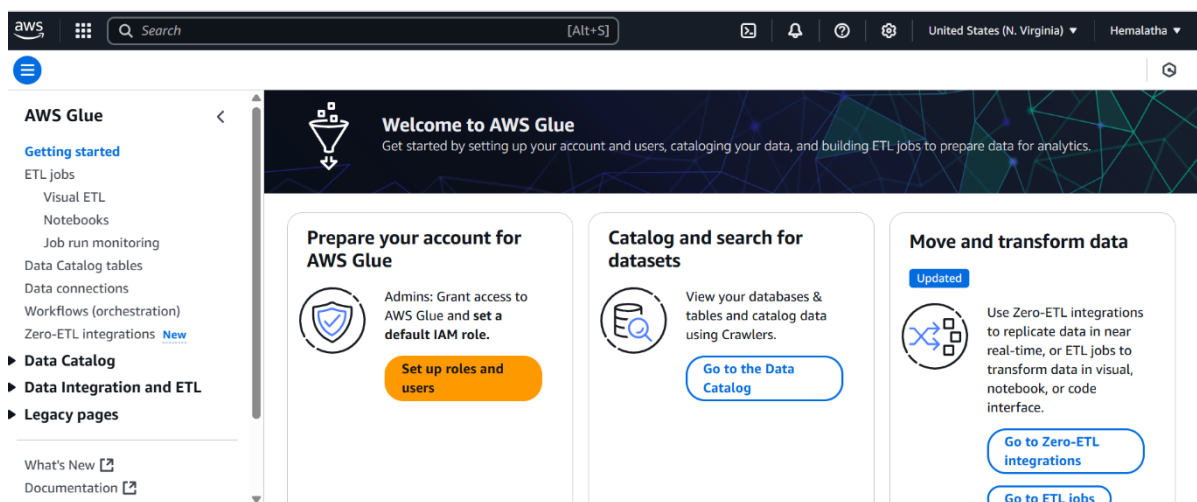
Common Use Cases

- 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.



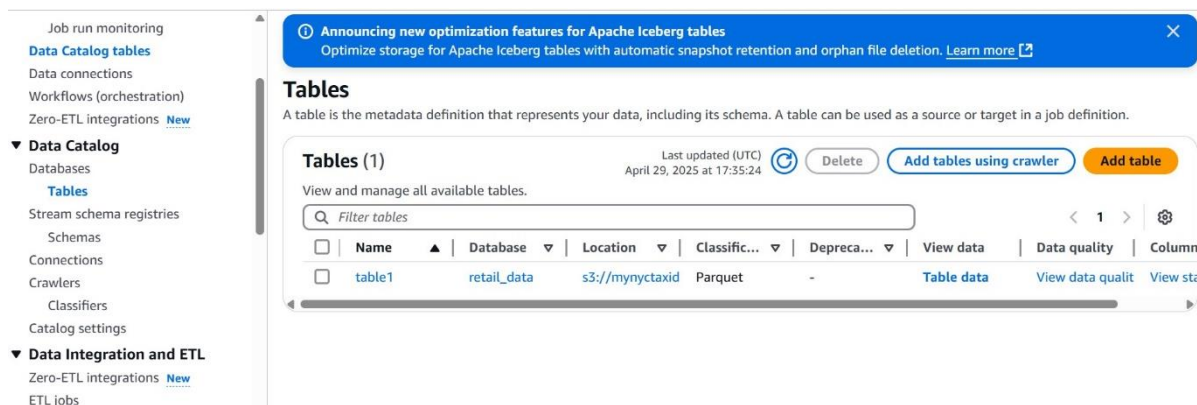
2. Create a New Database:

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



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 top screenshot shows the AWS IAM console interface for the 'readaccess' role. The left sidebar contains the 'Identity and Access Management (IAM)' menu with options like Dashboard, Access management, User groups, Users, Roles, Policies, Identity providers, Account settings, and Root access management. The main content area shows the role details for 'arn:aws:iam::239710306715:role/readaccess', including the date 'April 29, 2025, 13:08 (UTC-04:00)', 'Last activity', and 'Maximum session duration' of 1 hour. The 'Permissions' tab is selected, showing 'Permissions policies (1)' and a list of policies including 'AmazonS3ReadOnlyAccess'.

The bottom screenshot shows the AWS IAM console interface for the 'Roles' page. A green notification banner at the top states 'Role readaccess created.' with a 'View role' button. The left sidebar is the same as the top screenshot. The main content area shows the 'Roles Anywhere' section, which includes a 'Manage' button and three cards: 'Access AWS from your non AWS workloads', 'X.509 Standard', and '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 AWS Glue console with the 'Add crawler' wizard. The left sidebar lists navigation options: Getting started, ETL jobs, Visual ETL, Notebooks, Job run monitoring, Data Catalog tables, Data connections, Workflows (orchestration), Zero-ETL integrations, Data Catalog (expanded), Databases, Tables, Stream schema registries, Schemas, Connections, Crawlers (selected), and Classifiers. The main area is titled 'Set crawler properties' and includes a progress bar with five steps: Step 1 (Set crawler properties), Step 2 (Choose data sources and classifiers), Step 3 (Configure security settings), Step 4 (Set output and scheduling), and Step 5 (Review and create). The 'Crawler details' section has a 'Name' field with the value 'sales_data_crawler' and a 'Description' field with the placeholder 'Enter a description'. Below this is a 'Tags' section with a note: 'Use tags to organize and identify your resources.'

This screenshot shows the 'Add data source' modal window. It prompts the user to 'Choose the source of data to be crawled.' with a dropdown menu set to 'S3'. Below this, the 'Network connection - optional' section has a dropdown and an 'Add new connection' button. The 'Location of S3 data' section has two radio buttons: 'In this account' (selected) and 'In a different account'. The 'S3 path' section includes a text input with 's3://mynuctaxidata', a 'View' button, and a 'Browse S3' button. At the bottom, there are 'Cancel' and 'Add an S3 data source' buttons.

The screenshot displays the 'Configure security settings' step of the 'Add crawler' wizard. The progress bar shows Step 3 is active. The 'IAM role' section has a dropdown for 'Existing IAM role' set to 'readaccess', with 'Create new IAM role' and 'Update chosen IAM role' buttons. Below this is a 'Lake Formation configuration - optional' section with a checkbox 'Use Lake Formation credentials for crawling S3 data source' which is currently unchecked. The bottom of the screen shows the start of the 'Security configuration - optional' section.

aws | glue

AWS Glue > Crawlers > Add crawler

Getting started
ETL jobs
Visual ETL
Notebooks
Job run monitoring
Data Catalog tables
Data connections
Workflows (orchestration)
Zero-ETL integrations **New**

▼ Data Catalog
Databases
Tables
Stream schema registries
Schemas
Connections
Crawlers
Classifiers
Catalog settings

Step 2
● Choose data sources and classifiers
Step 3
● Configure security settings
Step 4
● **Set output and scheduling**
Step 5
○ Review and create

Output configuration Info

Target database
retail_data

Clear selection Add database

Table name prefix - optional
Type a prefix added to table names

Maximum table threshold - optional
This field sets the maximum number of tables the crawler is allowed to generate. In the event that this number is surpassed, the crawl will fail with an error. If not set, the crawler will automatically generate the number of tables depending on the data schema.
Type a number greater than 0

► Advanced options

Crawler schedule

aws | glue

AWS Glue > Crawlers > sales_data_crawler

Getting started
ETL jobs
Visual ETL
Notebooks
Job run monitoring
Data Catalog tables
Data connections
Workflows (orchestration)
Zero-ETL integrations **New**

▼ Data Catalog
Databases
Tables
Stream schema registries
Schemas
Connections
Crawlers
Classifiers
Catalog settings

One crawler successfully created
The following crawler is now created: "sales_data_crawler"

sales_data_crawler
Last updated (UTC)
April 29, 2025 at 17:20:22

Run crawler Edit Delete

Crawler properties

Name sales_data_crawler	IAM role readaccess	Database retail_data	State READY
Description -	Security configuration -	Lake Formation configuration -	Table prefix -
Maximum table threshold -			

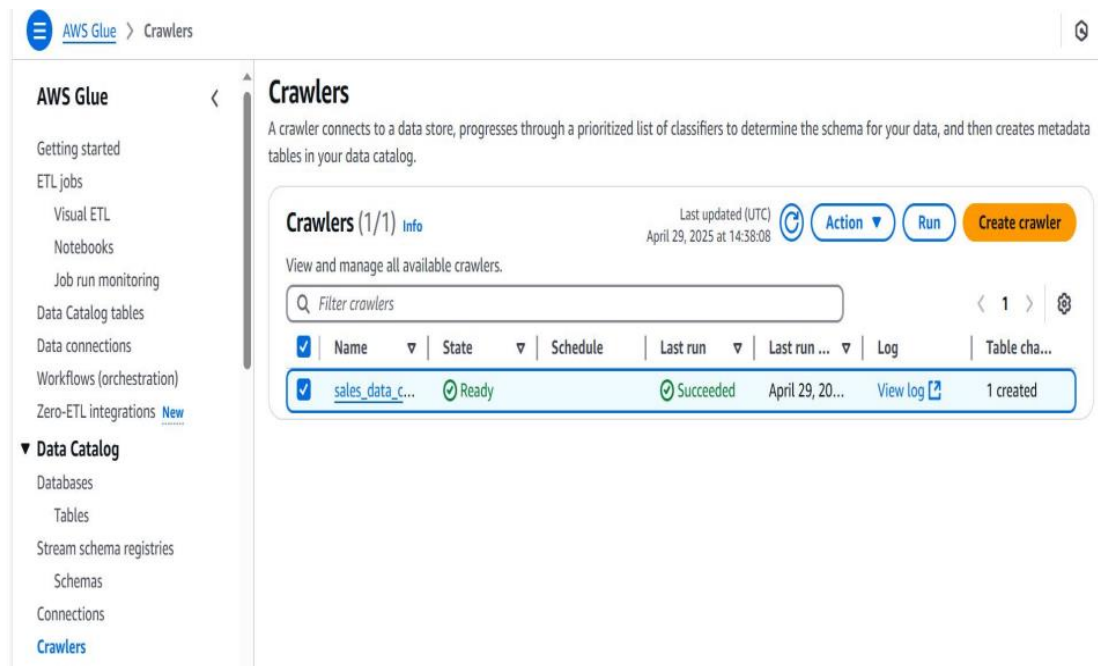
► Advanced settings

Crawler runs | Schedule | Data sources | Classifiers | Tags

5. Run the Crawler:

Select the crawler and click Run Crawler.

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



6. Verify the Metadata:

Go to the Tables section in the Data Catalog.

Review the schema — columns, data types, and partition

