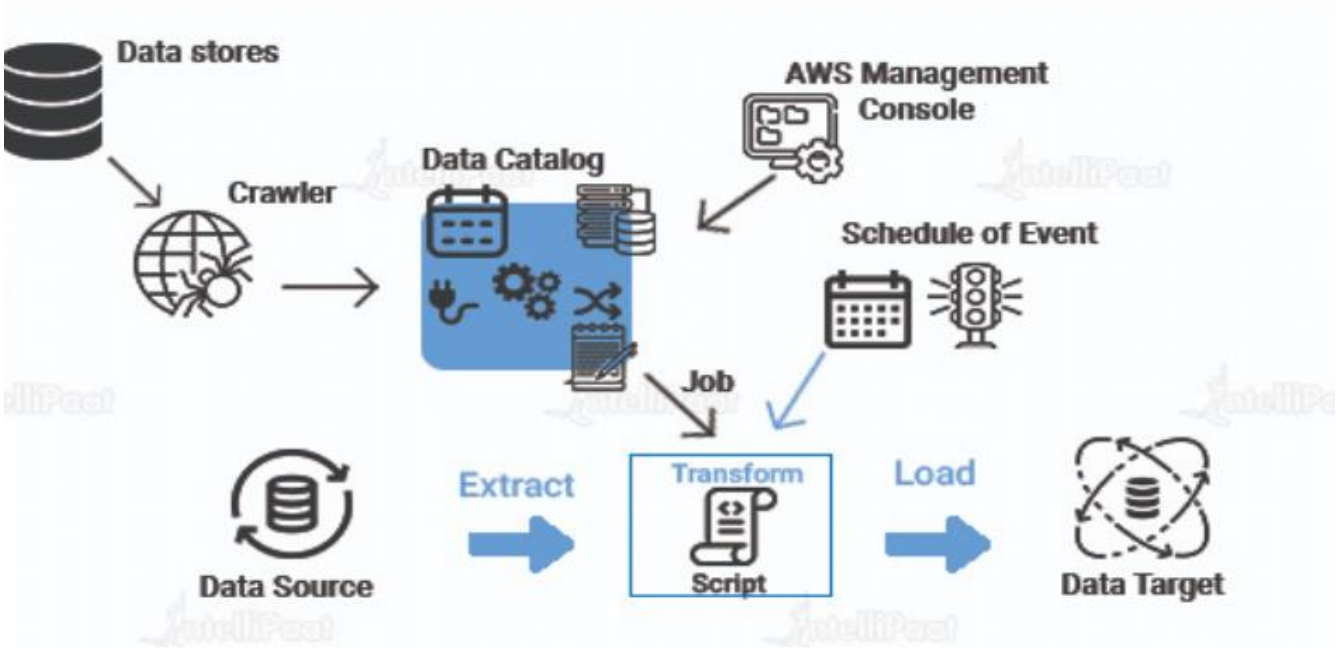
AWS GLUE

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.

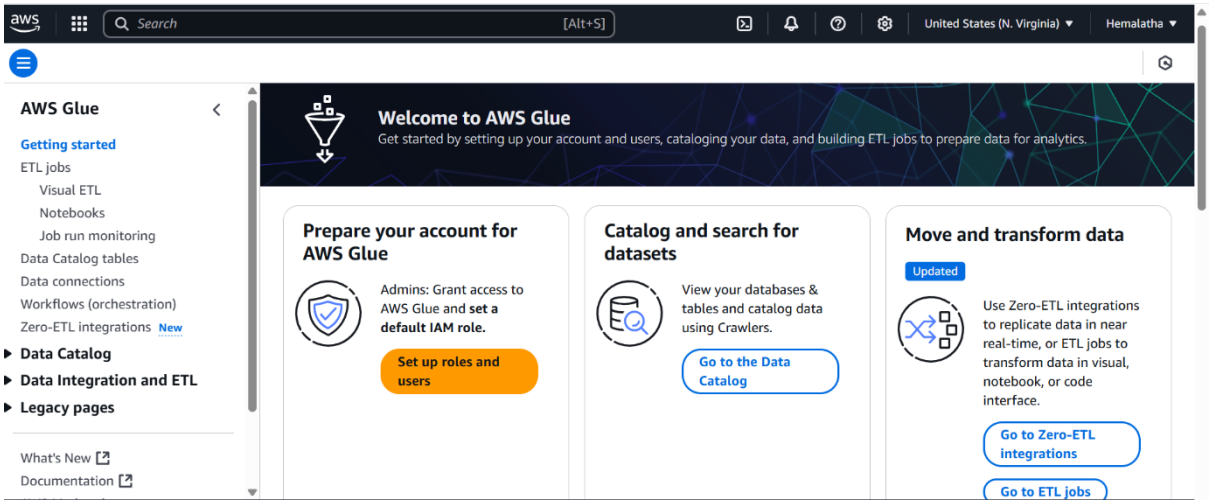
Common Use Cases

- 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.



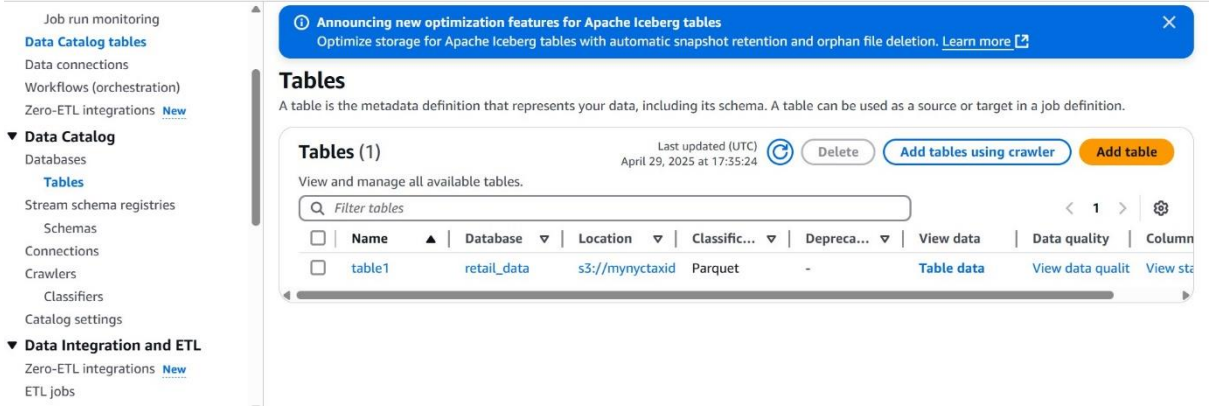
2. Create a New Database:

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



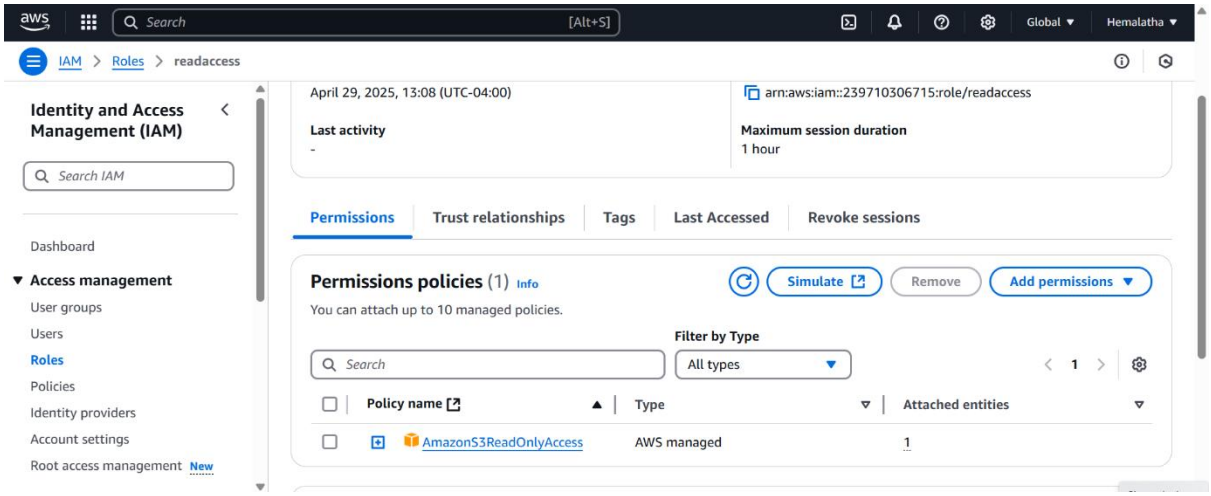
3. Set Up IAM Role:

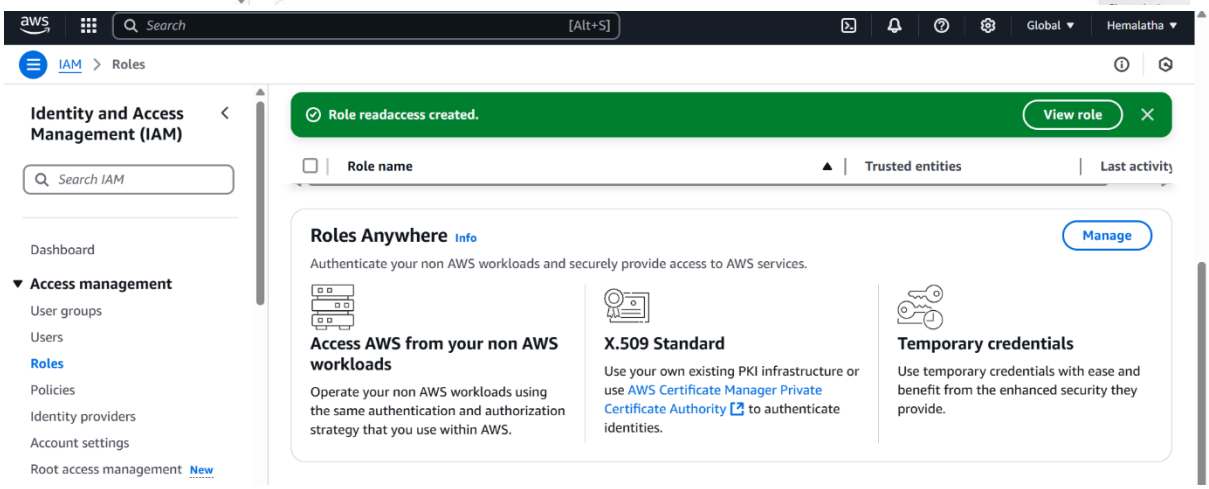
Ensure a role exists with these policies:

AWS Glue Service Role

AmazonS3ReadOnlyAccess

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





4. Create a Crawler:

Go to Crawlers → Add Crawler.

Name: sales_data_crawler

Data Store: S3

Path: s3://retail-data-bucket/sales_data/

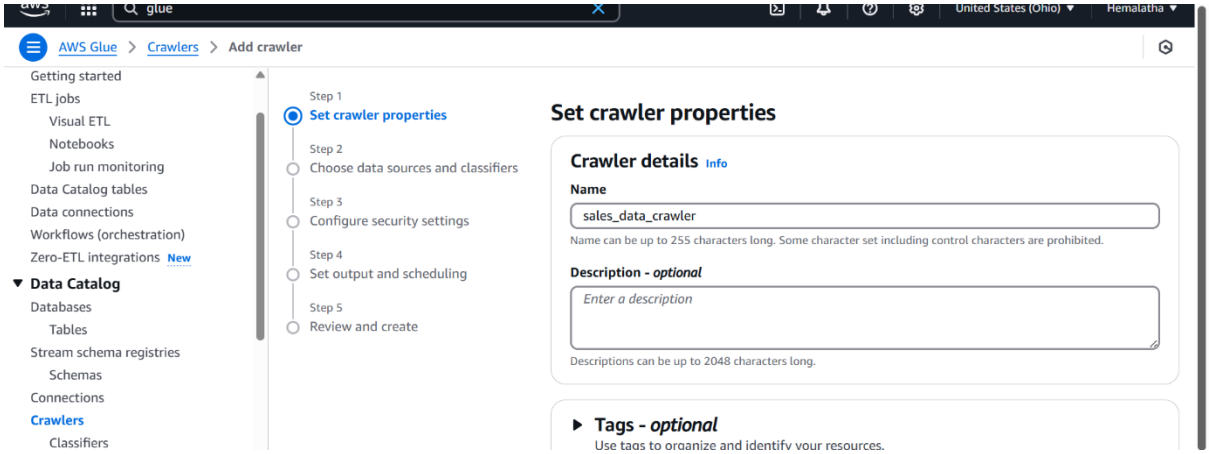
IAM Role: Use the role created earlier.

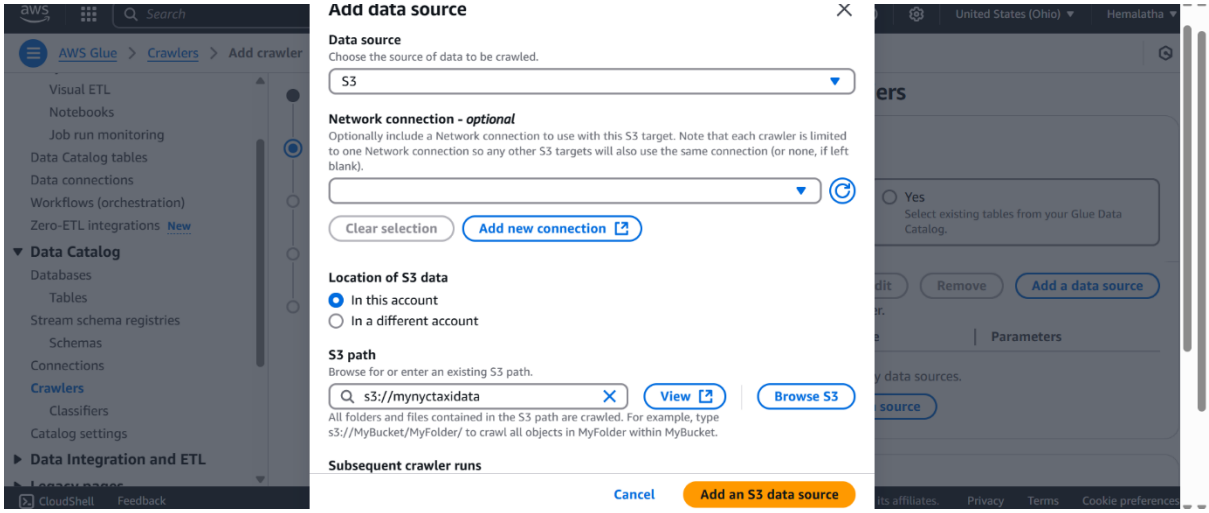
Output Database: Select retail_sales_db

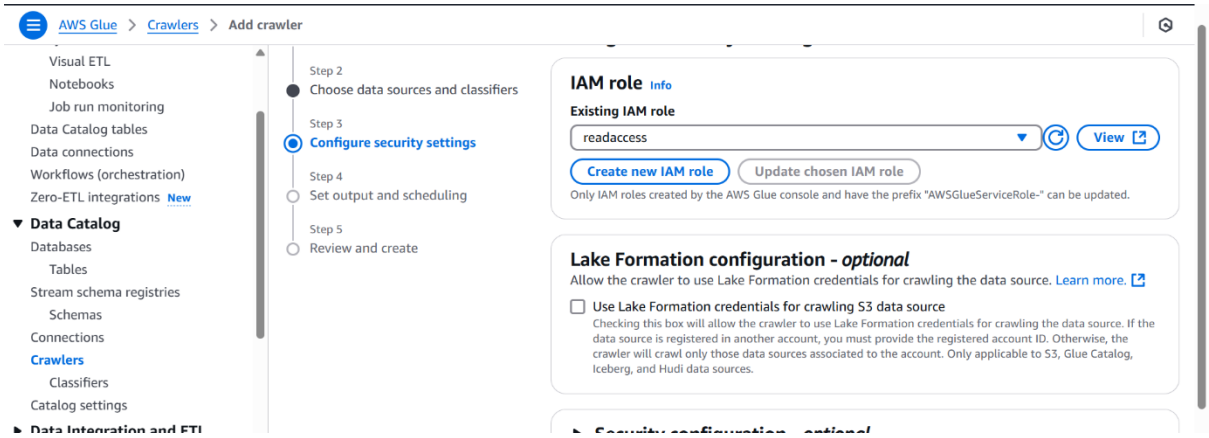
Run Frequency: Run on demand (for now)

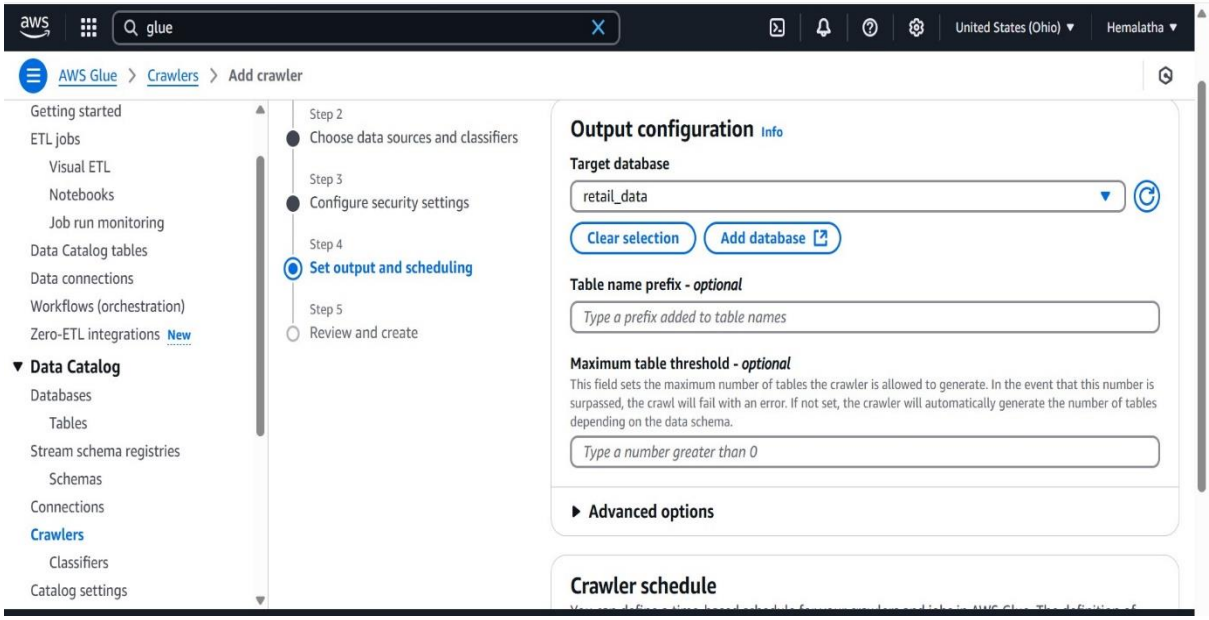
Review and create the crawler.

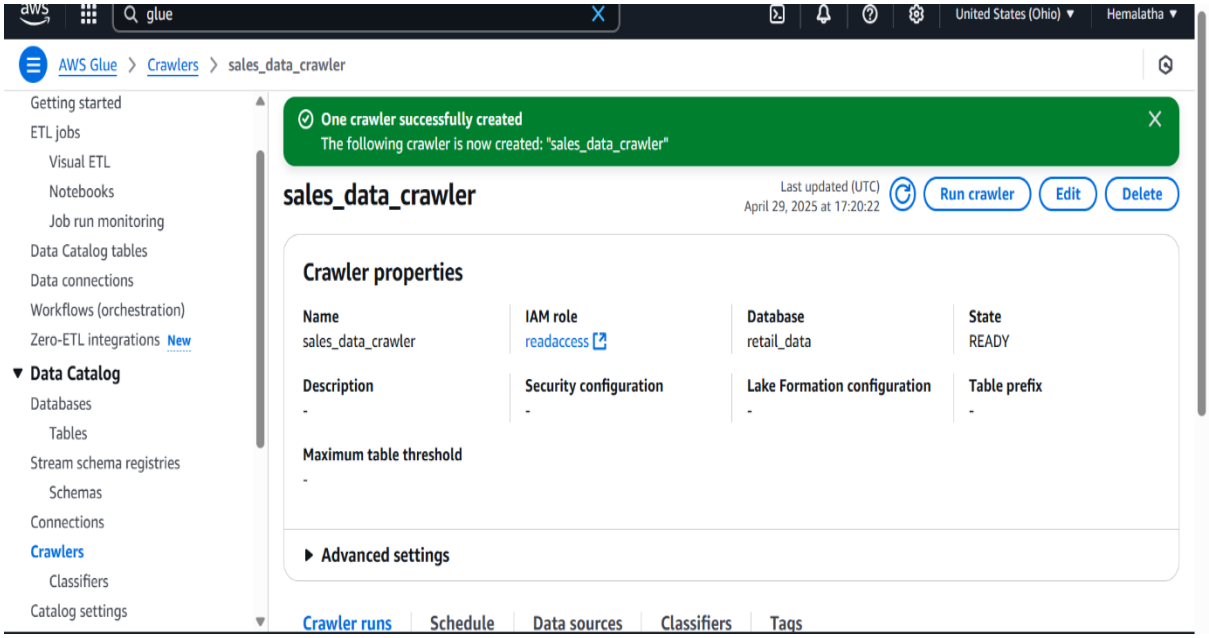
Open AWS glue → data catalog → crawlers →create crawler







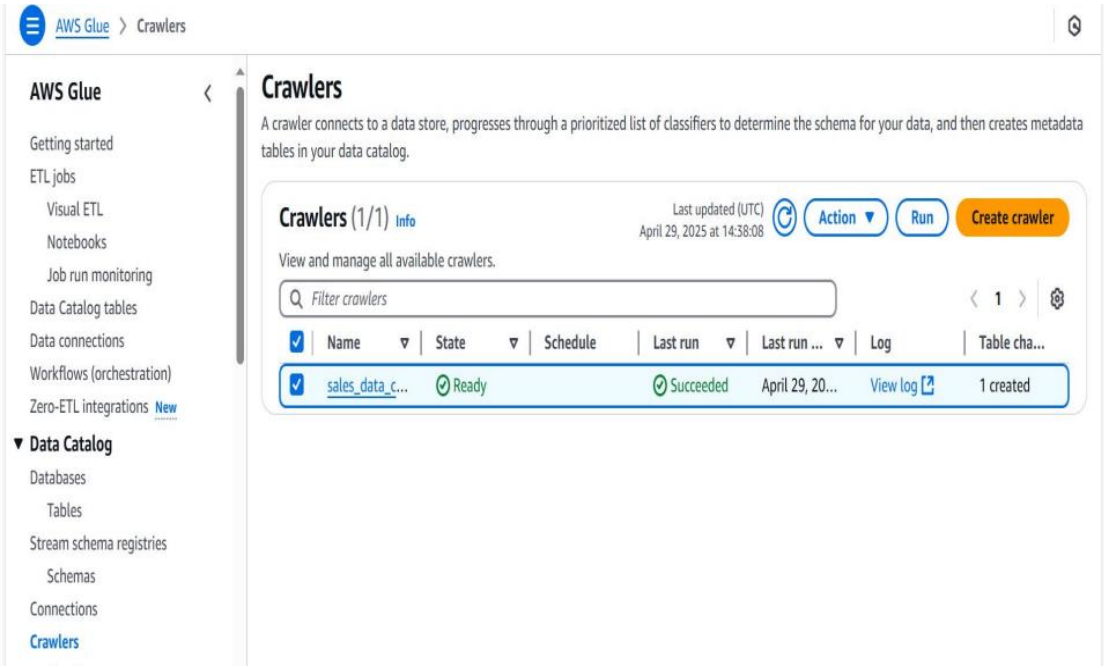




5. Run the Crawler:

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