Step Of AWS Projects :: Create a Data Pipeline with AWS Glue & Athena

**🎯 Objective:** Build a **serverless ETL pipeline** using **AWS Glue** to process and analyze CSV data stored in **S3** with **Athena**.

**Step 1: Upload CSV Data to S3**

**Why?** AWS Glue reads data from S3, so we need a dataset stored in an S3 bucket.

**1.1. Create an S3 Bucket**

1. Go to **AWS Console** → **S3** → Click **Create Bucket**
2. Enter **Bucket Name**: my-data-pipeline-bucket
3. Choose a region (e.g., us-east-1)
4. Leave everything **default** and click **Create Bucket**

**1.2. Upload Sample CSV File**

1. Click on the newly created bucket
2. Click **Upload** → **Add Files**
3. Upload a CSV file (e.g., sales\_data.csv)
4. Click **Upload**

**📌 Sample CSV Data (sales\_data.csv):**

**Step 2: Create an AWS Glue Crawler**

**Why?** The **AWS Glue Crawler** scans our S3 data and automatically creates a **table schema** in Glue Data Catalog.

**2.1. Open AWS Glue**

1. Go to **AWS Console** → **AWS Glue**
2. Click **Crawlers** → Click **Create Crawler**

**2.2. Configure Crawler**

1. **Crawler Name:** sales-data-crawler
2. **Data Source:** Choose **S3**
3. **Include Path:** Select s3://my-data-pipeline-bucket/
4. **IAM Role:** Create a new IAM role with AWSGlueServiceRole
5. **Output Database:** Create a new Glue database (sales\_db)
6. **Run Frequency:** Choose **Run on Demand**

**2.3. Run the Crawler**

1. Click **Run Crawler**
2. Wait for it to complete (takes 2–5 minutes)
3. Check **AWS Glue → Tables** → Your dataset should be available! 🎉

**Step 3: Query Data with AWS Athena**

**Why?** Athena allows **SQL queries** on our S3 data without setting up a database.

**3.1. Open AWS Athena**

1. Go to **AWS Console** → **Athena**
2. Click **Get Started** (if first time)
3. In **Settings**, set the **Query Result Location**:
4.  In Athena, select the **sales\_db** database
5.  Run this SQL query to check the data:

**Step 4: Automate Data Processing (Optional)**

**Want to process and transform data?** Use AWS Glue ETL jobs:

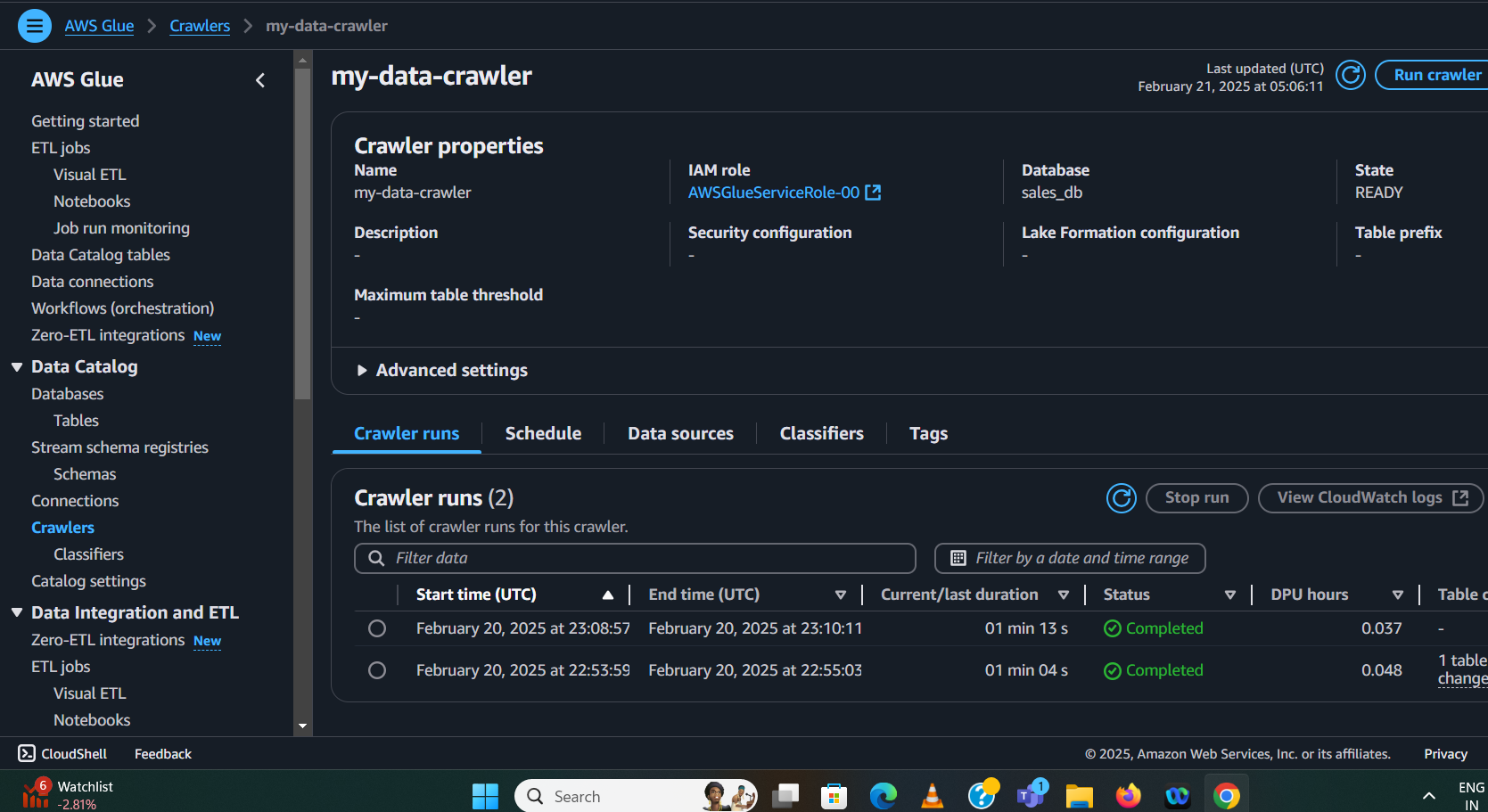
**4.1. Create an AWS Glue ETL Job**

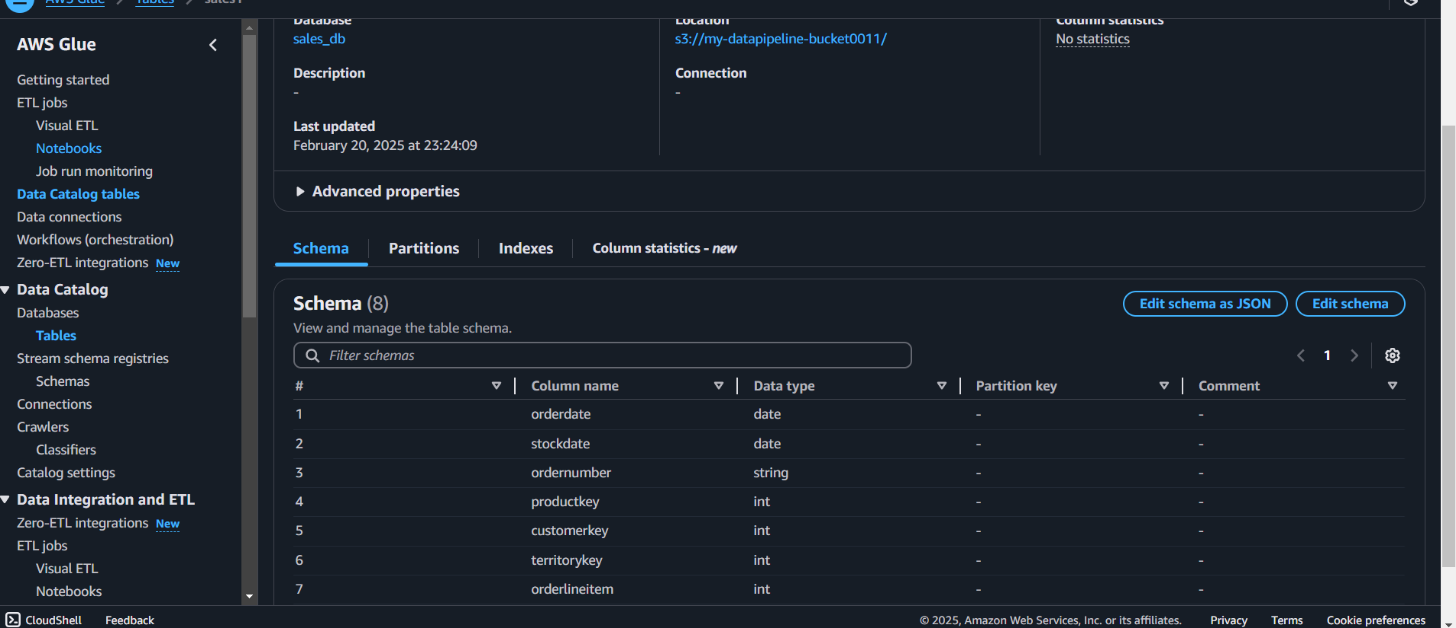
1. Go to **AWS Glue** → **Jobs** → Click **Create Job**
2. Choose **"Visual with a source and target"**
3. **Source:** Select the Glue table (sales\_data)
4. **Transformations:** (Optional) Add column renaming, filtering, etc.
5. **Target:** Choose **S3 (Parquet Format)**
6. Click **Save and Run**

🎉 Now, your data pipeline is **automated!**

**Images of working projects**

**First images :---------------------------------------------------------------------------------------------**



Second Images : ------------------------------------------------------------------------------------------

Third images :---------------------------------------------------------------------------------------------

