Project Title:

Real-Time and Historical Sales Data Integration Using Amazon Redshift Spectrum

---

Problem Scenario:

An e-commerce company's data engineering team needs to analyze both **real-time sales data stored in Amazon Redshift** and **historical sales data stored in Amazon S3** (in Parquet/CSV format). The challenge is to query both datasets **together** without:

- Duplicating the S3 data into Redshift
- Manually importing historical files
- Running frequent ETL pipelines

Objective:

To build a solution that enables seamless querying of **transactional (live)** and **historical (S3)** sales data using **Amazon Redshift Spectrum**.

Solution Overview:

We implemented a hybrid analytics system using:

- 🗃️ **Amazon S3** for storing historical data partitioned by `year` and `month`
- 🧠 **AWS Glue Data Catalog** for crawling and cataloging external S3 data
- 🔴 **Amazon Redshift Serverless** for real-time querying
- 🌐 **Redshift Spectrum** to federate queries across internal Redshift and external S3 data

---

Implementation Steps:

1 Created and Uploaded Historical Sales Data to S3

- Created structured CSV files with fields:
    - `product_id`, `quantity`, `revenue`
    - (Partition columns `year` and `month` were *excluded* from CSV body but added as folder structure)

- Uploaded files using path structure:

  `s3://ecommerce-sales--data--bucket/historical_sales/year=2023/month=1/`

2 Configured AWS Glue Data Catalog

- Created Glue Database: `ecommerce_sales_db`
- Created a **Glue Crawler** pointing to:

  `s3://ecommerce-sales--data--bucket/historical_sales/`
- Crawler detected partitions and created a table:

  `ecommerce_sales__data__bucket`

  with inferred schema and partitions `year`, `month`

3 Created an IAM Role for Redshift Spectrum Access

- IAM Role Name: `my-iam-redshift-role`
- Attached the following policies:
  - `AmazonS3ReadOnlyAccess`
  - `AWSGlueConsoleFullAccess` (or `AWSGlueCatalogReadOnlyAccess`)
- Attached the IAM role to Redshift Serverless **namespace settings**

4 Created External Schema in Redshift

Connected Redshift to the Glue Data Catalog using:

```
CREATE EXTERNAL SCHEMA spectrum_schema
FROM DATA CATALOG
DATABASE 'ecommerce_sales_db'
IAM_ROLE 'arn:aws:iam::535002864770:role/my-iam-redshift-role'
CREATE EXTERNAL DATABASE IF NOT EXISTS;
```

5 Created Redshift Internal Table (Transactional Data)

```
CREATE TABLE redshift_sales (
  product_id INT,
  quantity INT,
  revenue FLOAT
);

-- Sample data
INSERT INTO redshift_sales VALUES (101, 2, 200), (102, 1, 120);
```

6 Joined Redshift and Spectrum Tables

Combined real-time and historical data using:

```sql
SELECT
  r.product_id,
  r.quantity AS live_qty,
  h.quantity AS historical_qty,
  r.revenue AS live_revenue,
  h.revenue AS historical_revenue
FROM
  redshift_sales r
JOIN
  spectrum_schema.ecommerce_sales__data__bucket h
ON
  r.product_id = h.product_id
WHERE
  h.year = 2023 AND h.month = 1;
```

This query successfully fetched data from both Redshift and S3 in a single result set.

Outcome:

- ✅ Successfully queried **live transactional** data and **historical S3** data together
- ✅ Verified **Redshift Spectrum** setup and IAM access
- ✅ Achieved **zero duplication** and no data import overhead

---

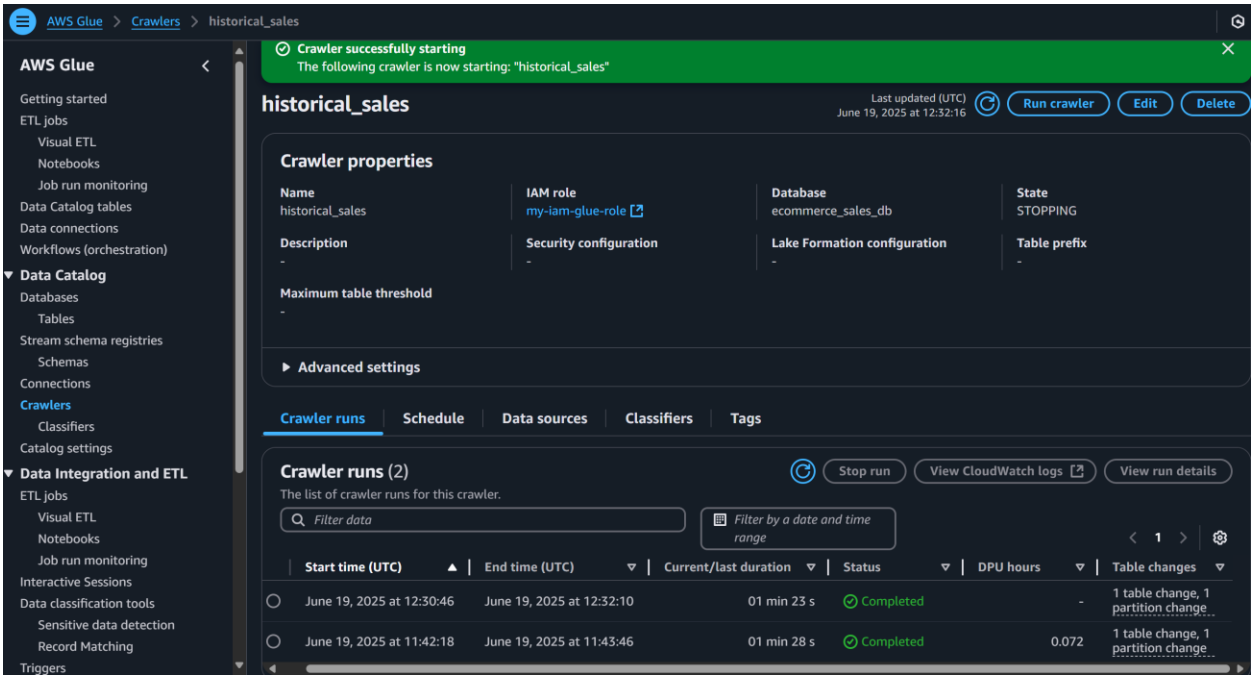
Benefits:

- 🔄 No need to load S3 data into Redshift
- 📊 Efficient partition-based pruning from S3
- ⚡ Fast, federated query across real-time and historical sales
- 🧩 Unified analytics layer for dashboards and reports

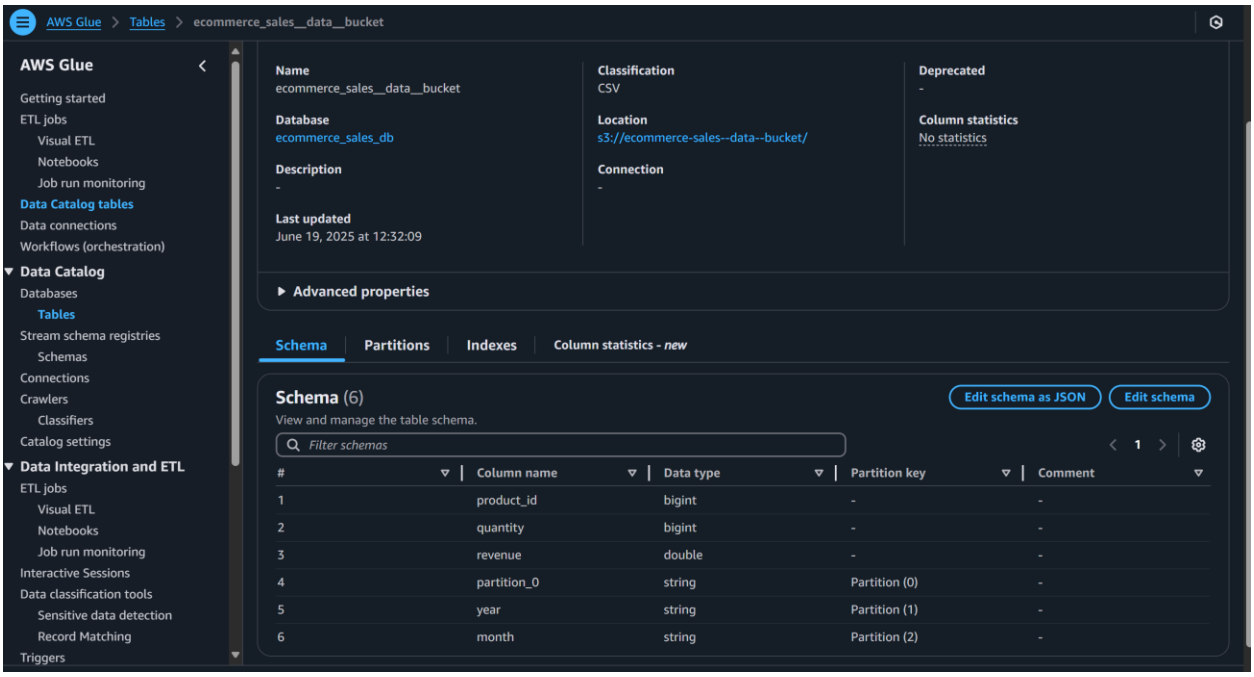
Workspace Output (Paste Screenshots/Logs Below):

`[S3 folder structure screenshot]`

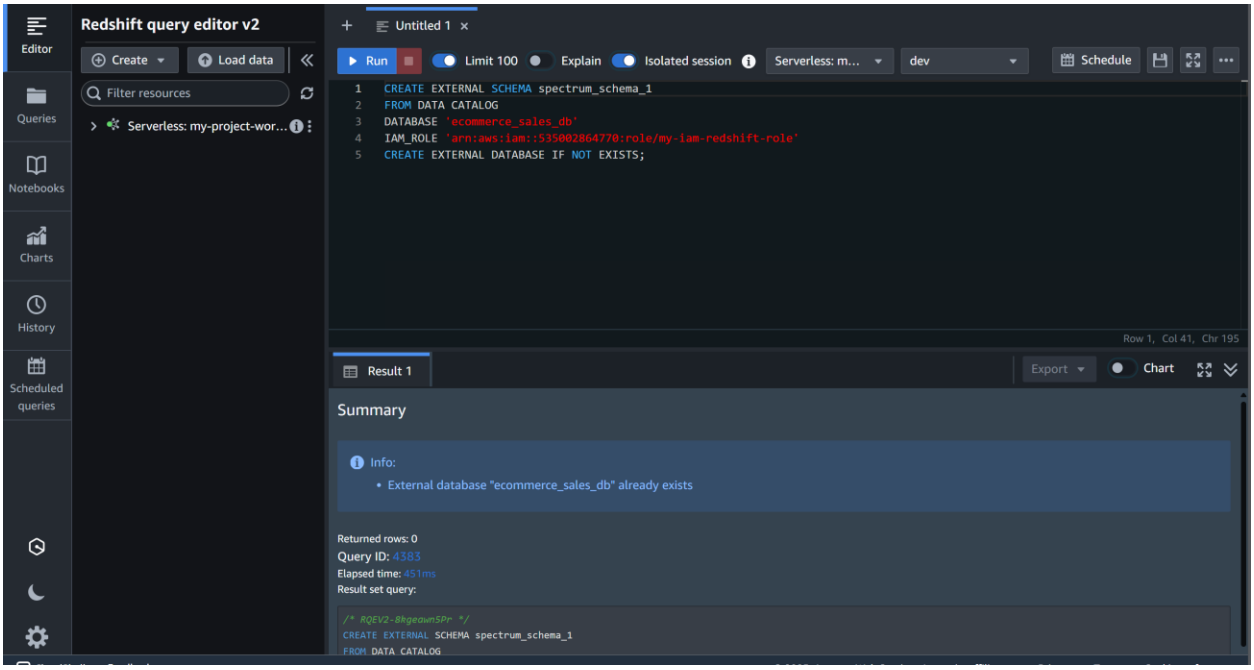
[Glue crawler setup screenshot]



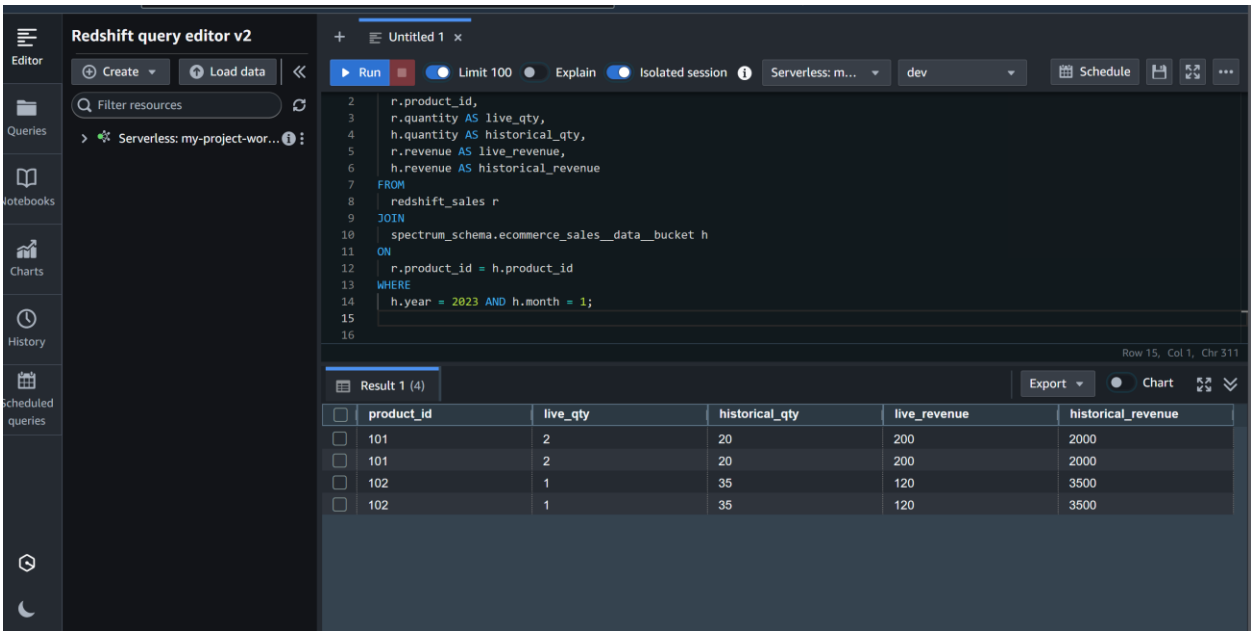
[Glue catalog table structure]



[Redshift SQL editor queries and output]



[Query result screenshot combining Redshift + Spectrum data]



Prepared by:

Atmakuru Siva Sandeep

Date: 19 June 2025

Tags:

#AWS #RedshiftSpectrum #GlueCatalog #S3Analytics #HybridDataWarehouse #FederatedQueries