# Batch Time Analysis of Transactional Data

*Christina Grys*

Lenodo is a multinational e-commerce organization that sells products directly to consumers. The database administrator exports the data every night in a CSV file, but this export functionality is unused. Lenodo wants to use this data to uncover insights about the most-sold item and the countries where customers have bought this item.

You are a data analytics consultant, and you're asked to provide valuable insights and statistics across products, brands, categories, segments to the marketing, product, sales, and procurement teams and inform them about which product has the highest amount of sales and which product and its marketing needs the most improvement. These statistics will help to run effective digital marketing campaigns. The scope of this project is limited to data engineering and analysis.

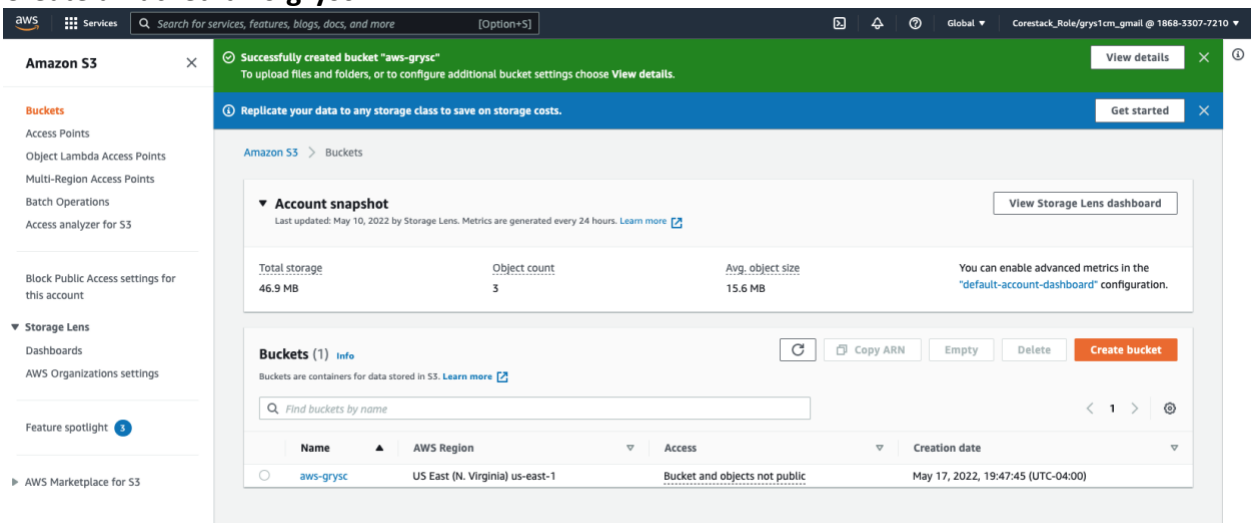
**Objective:**

To use AWS Big Data stack for data engineering to analyze transactions, uncover patterns, and share actionable insights

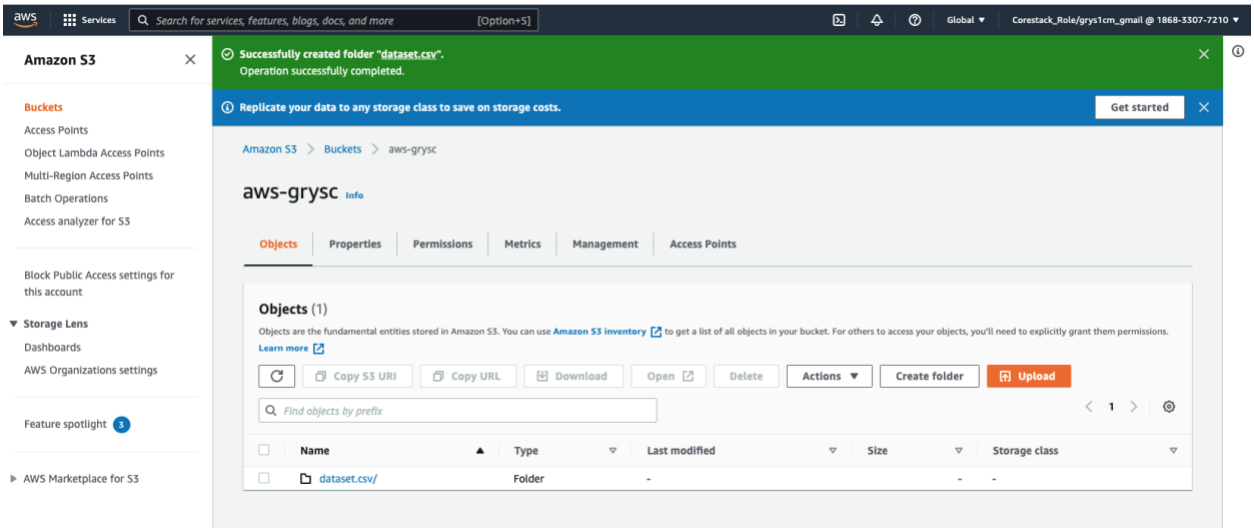
**Steps to perform:**

1. Create an S3 bucket with a unique name and upload the CSV file to the S3 bucket (ensure that the file is in UTF-8 format only)
2. Create a crawler to crawl the CSV data and generate a metadata catalog
3. Create a Glue job to transform the data into the Parquet format as CSV is not optimal for data warehouse queries
4. Add another crawler to crawl the Parquet data files to generate the metadata catalog of the Parquet file in order to query it with Athena
5. Query the data to identify the best-selling item and countries where customers have bought the most-sold item using Athena

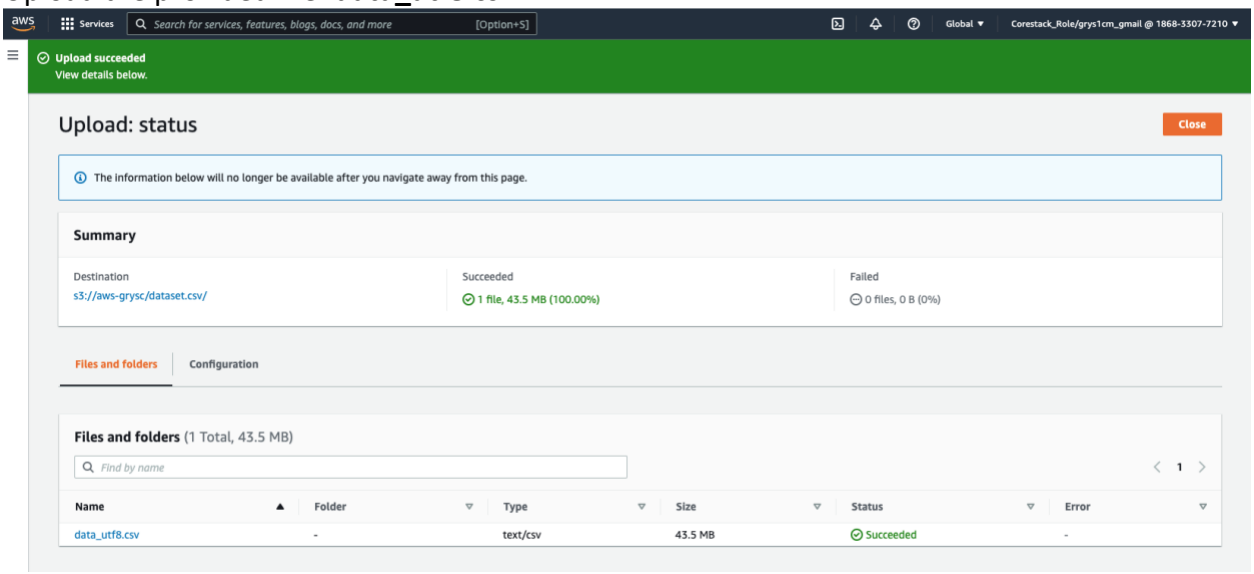
## Create a Bucket: **aws-grysc**



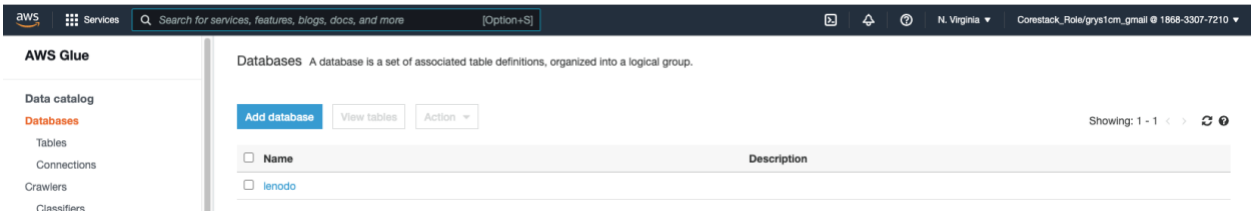
Create A Folder: **dataset.csv**



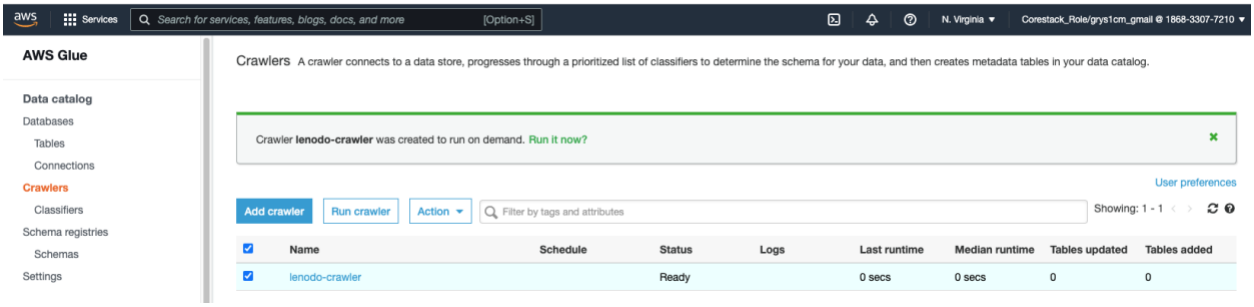

Upload the provided file: **data_utf8.csv**




Add Database: **lenodo**



Add Crawler: lenodo-crawler





Crawlers > lenodo-crawler

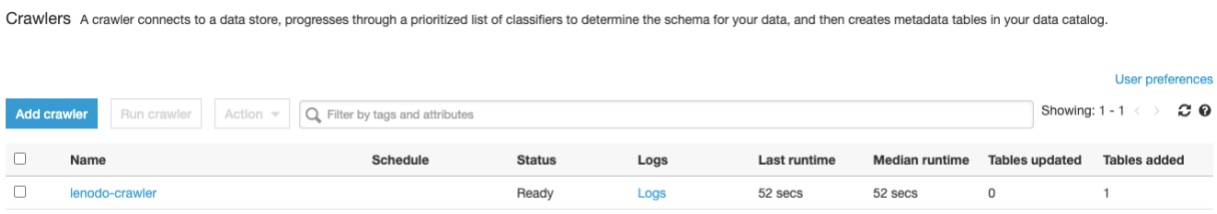
Run crawler | Edit

| | |
|---|---|
| **Name** | lenodo-crawler |
| **Description** | |
| **Create a single schema for each S3 path** | false |
| **Security configuration** | |
| **Tags** | - |
| **State** | Ready |
| **Schedule** | |
| **Last updated** | Tue May 17 19:55:52 GMT-400 2022 |
| **Date created** | Tue May 17 19:55:52 GMT-400 2022 |
| **Database** | lenodo |
| **Table level** | |
| **Service role** | service-role/AWSGlueServiceRole-lenodo-project |
| **Selected classifiers** | |
| **Data store** | S3 |
| **Include path** | s3://aws-grysc/dataset.csv |
| **Connection** | |
| **Exclude patterns** | |

## Configuration options

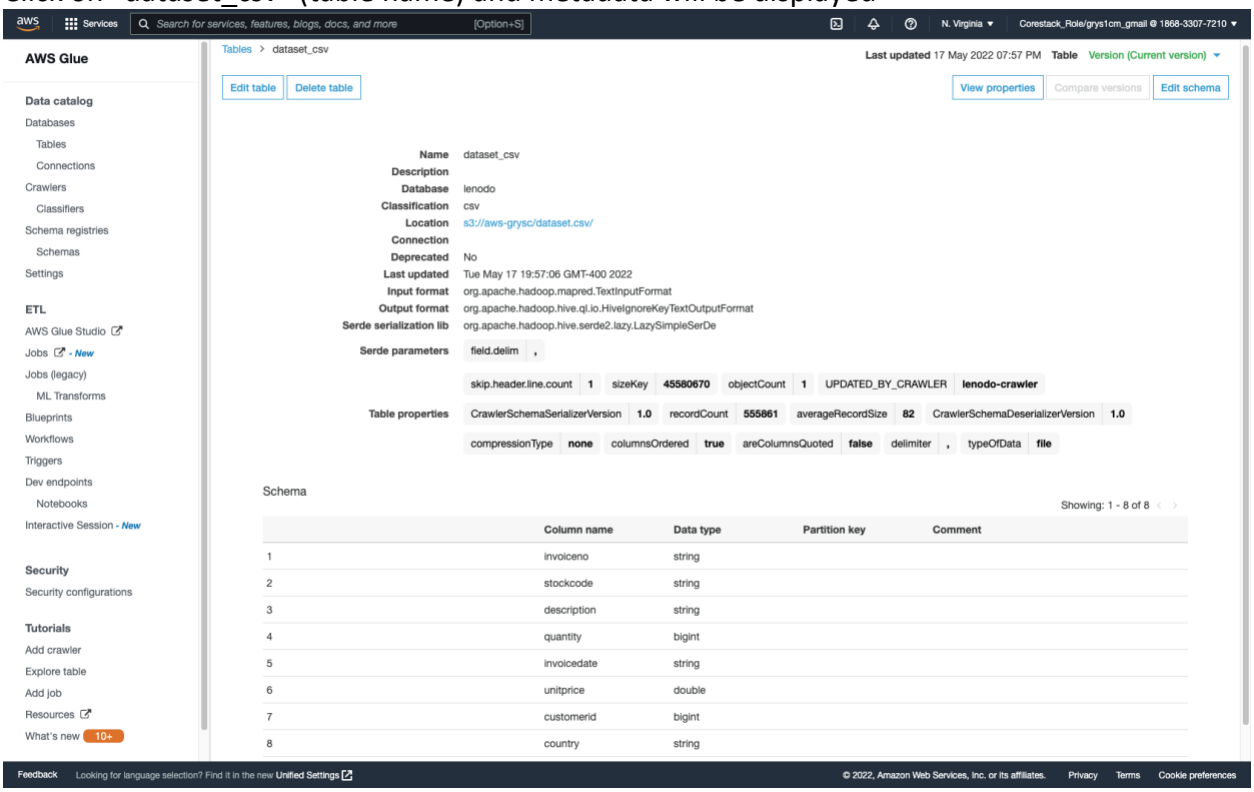
| | |
|---|---|
| **Schema updates in the data store** | Update the table definition in the data catalog. |
| **Object deletion in the data store** | Mark the table as deprecated in the data catalog. |

Run Crawler - Ensure the Table is added to “1”

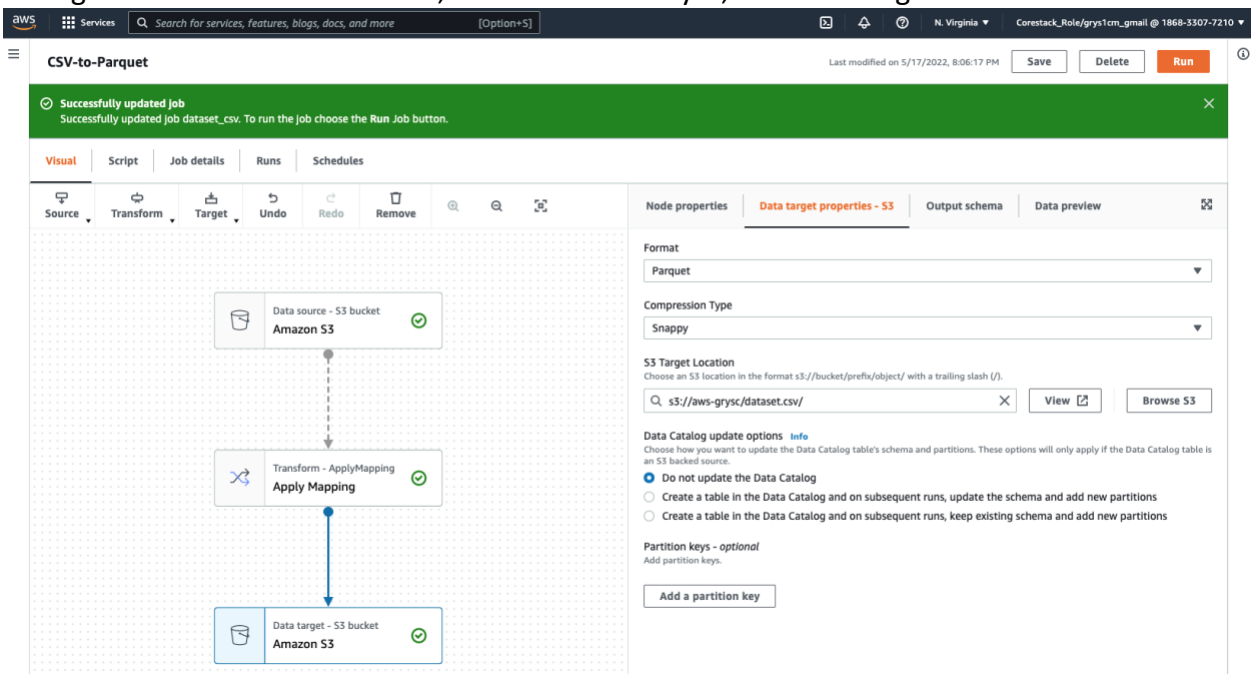



Check Table on the left panel bar under “Tables”

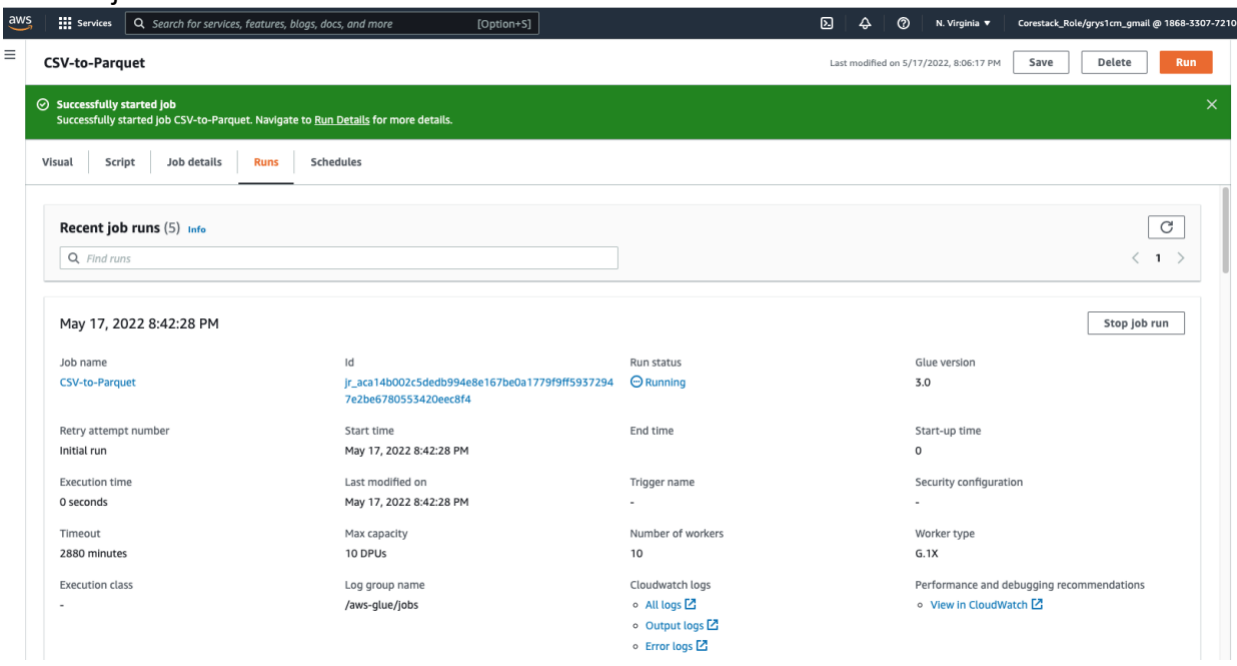

Click on “dataset_csv” (table name) and metadata will be displayed



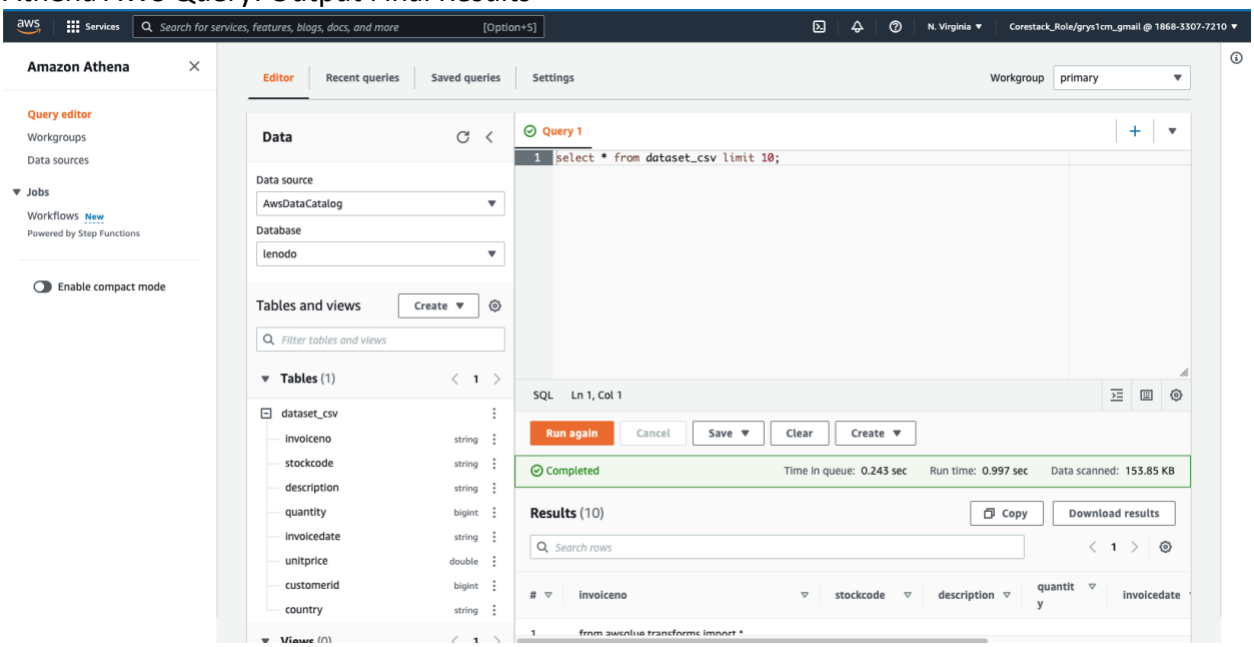
Configurations of the data source, transformation layer, and data target all added.



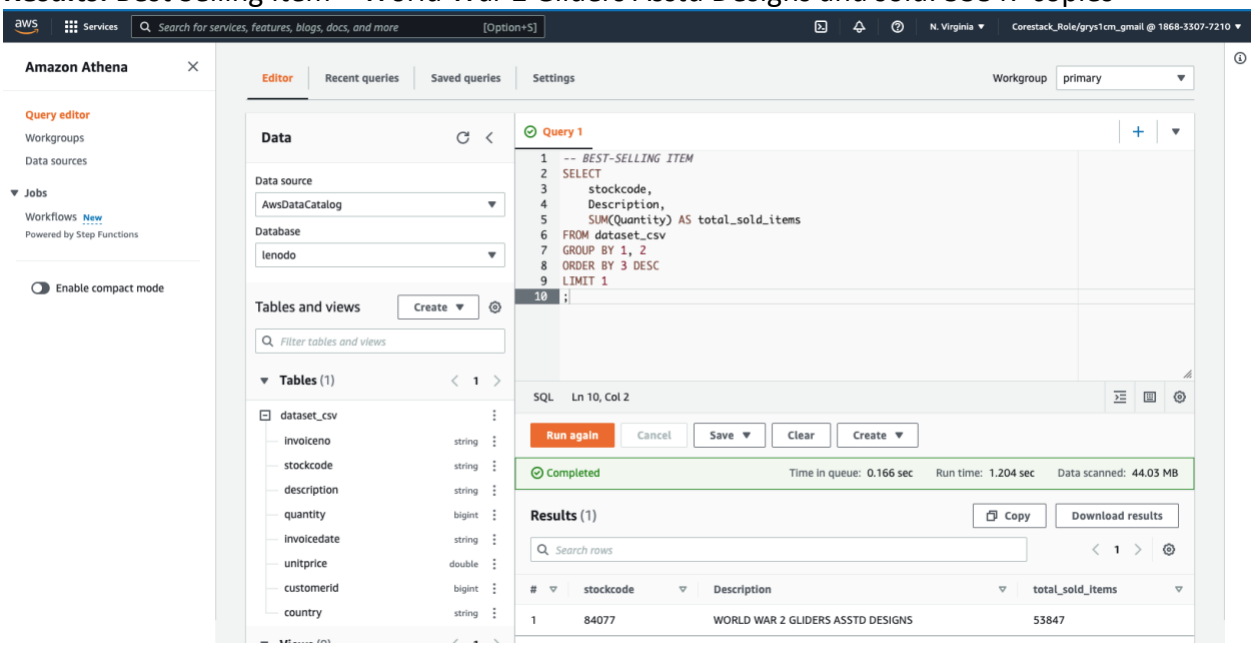

Run the job



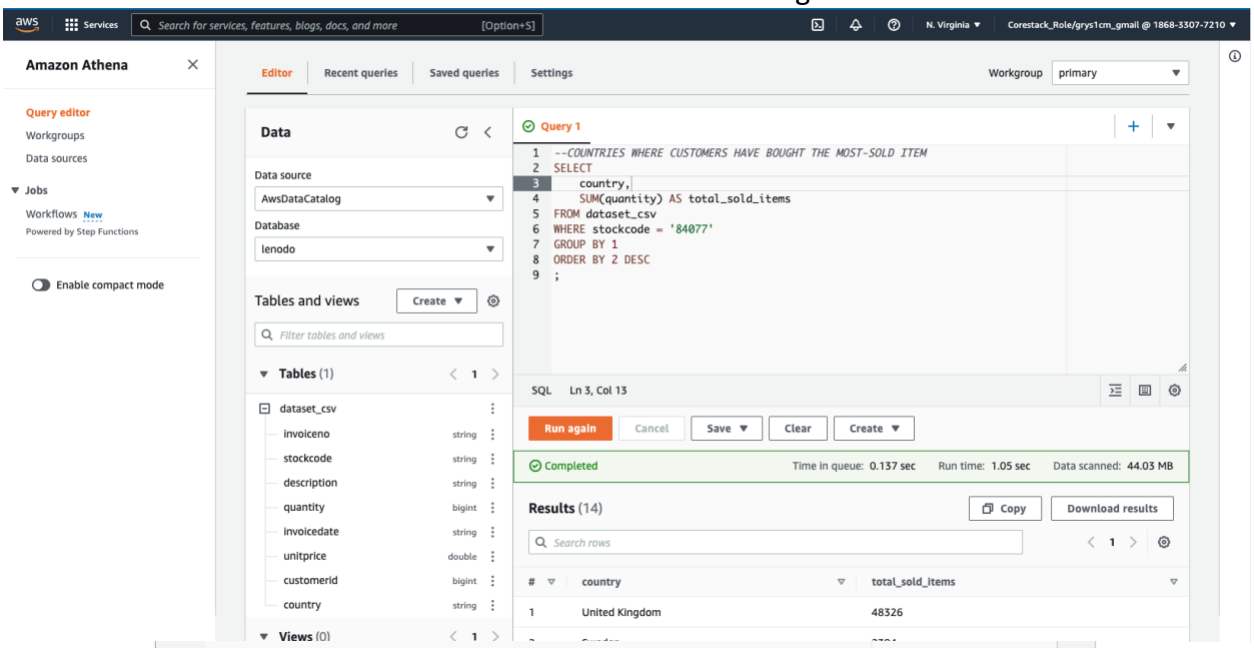
Athena AWS Query: Output Final Results




**Results:** Best Selling Item = World War 2 Gliders Asstd Designs and sold: 53847 copies



**Results** for Countries where customers have bought the most-sold item




SQL Ln 3, Col 13

Run again | Cancel | Save | Clear | Create

Completed — Time in queue: 0.137 sec — Run time: 1.05 sec — Data scanned: 44.03 MB

**Results** (14) — Copy | Download results

Search rows

| # | country | total_sold_items |
|---|---|---|
| 1 | United Kingdom | 48326 |
| 2 | Sweden | 2304 |
| 3 | EIRE | 816 |
| 4 | Japan | 577 |
| 5 | France | 528 |
| 6 | Spain | 288 |
| 7 | Canada | 288 |
| 8 | Denmark | 144 |
| 9 | Switzerland | 144 |
| 10 | Hong Kong | 144 |
| 11 | Germany | 96 |
| 12 | Unspecified | 96 |
| 13 | Portugal | 48 |
| 14 | Norway | 48 |