**PROJECT**

**Comparative Data Solutions  
[Using Azure Technologies]**

***Submitted by***

***Chinchu Thampi***

**Table of contents**

* Project overview
* Key Components
* Business Requirement
* Primary objective
* Schema description for the csv file Financials.csv (16 Columns)
* Dataset Overview

**Serverless SQL Pool**

* Create database, schema and external file format
* Bronze Layer: External Table Creation
* PIPELINE: Created a Stored Procedure for the bronze external table.
* Silver layer creation - Data Cleaning & basic transformations
* PIPELINE: Created a Stored Procedure for the silver external table.
* Creating Gold layer Table
* PIPELINE: Gold layer Table creation
* Company Financial sales Report (By profit)- Power BI Visualization

**Dedicated SQL Pool**

* Created external data source and file format
* Created a temporary external table (for staging)
* Automation using Pipeline-
* Created the bronze layer table
* Pipeline- Bronze table creation- Stored procedure creation
* Silver layer creation - Data Cleaning & basic transformations
* Pipeline - Silver table creation, Stored procedure creation
* Creating Gold layer Table
* Pipeline- Single stored procedure for all gold tables
* Project Summary, comparison and Conclusion.

**Scenario 3:Batch Processing for Large Structured Data**

**Project Overview**

Perform batch processing for financial data stored in Azure Data Lake Storage, by following the medallion architecture. The work includes two scenarios in synapse analytics Serverless SQL pool and the Dedicated SQL pool.

**Key Components**:

* Azure Synapse Dedicated SQL Pool: For large-scale batch processing and data warehousing.
* Azure Synapse Serverless SQL Pool: For querying raw data or intermediate results stored in ADLS on-demand.

**Business Requirement**

To build a data pipeline to ingest, transform and analyze Sales dataset through batch processing

● Raw Data Ingestion into ADLS Gen2 (Bronze Layer)

● Use Synapse serverless and Dedicated SQL Pool to load and process data.

○ Data Profiling & Cleaning (Silver Layer)

○ Data Modeling & Transformations (Gold Layer)

● Leverage Synapse Serverless SQL Pool to run ad-hoc queries and generate insights from processed or raw data stored in ADLS.

**Primary Objective**

* This project aimed to design a scalable and efficient analytics pipeline using Azure Synapse Analytics, with a comparative focus on Dedicated SQL Pool and Serverless SQL Pool.
* The objective was to ingest, clean, transform, and visualize retail sales financial data to uncover business **insights around product performance, geographical profitability, and time-based sales trends.**

**Schema description for the csv file Financials.csv (16 Columns)**

| **Column Name** | **Data Type** | **Description** |
| --- | --- | --- |
| Segment | VARCHAR | Market segment or customer group |
| Country | VARCHAR | Country where the sale occurred |
| Product | VARCHAR | Type of product sold |
| Discount Band | VARCHAR | Discount category applied to the sale |
| Units Sold | VARCHAR | Quantity of units sold (needs conversion) |
| Manufacturing Price | VARCHAR | Cost to manufacture one unit (needs conversion) |
| Sale Price | VARCHAR | Selling price per unit (needs conversion) |
| Gross Sales | VARCHAR | Total sales before any discounts |
| Discounts | VARCHAR | Discount amount applied |
| Sales | VARCHAR | Net sales after discounts |
| COGS | VARCHAR | Cost of goods sold |
| Profit | VARCHAR | Net profit from the sale |
| Date | DATETIME | Date of the transaction |
| Month Number | INT | Numeric month (1 = January, etc.) |
| Month Name | VARCHAR | Full name of the month |
| Year | INT | Year of the transaction |

**The dataset looks like as follows:**

A screenshot of a computer

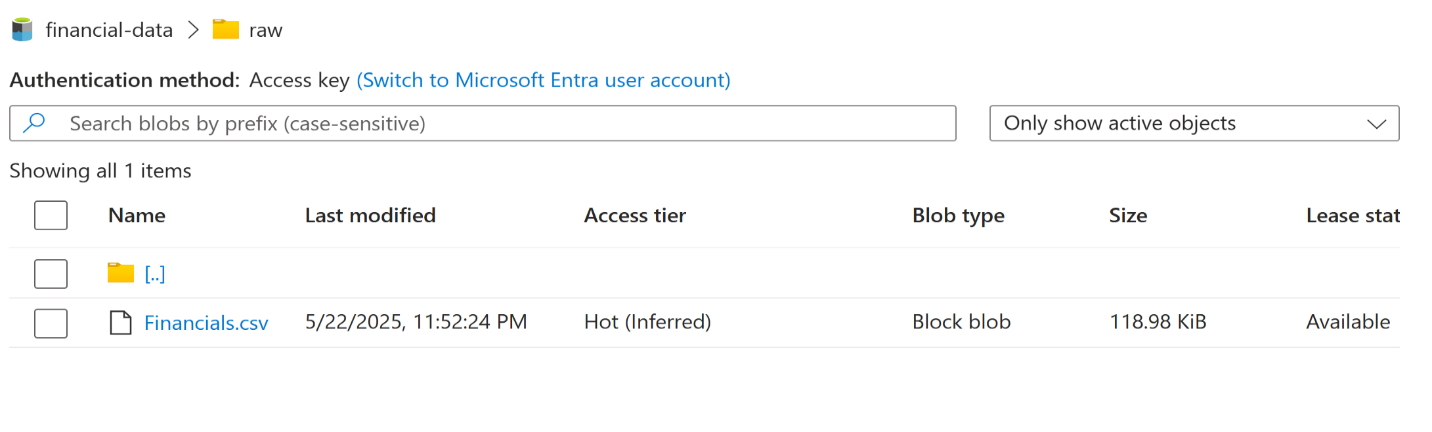
AI-generated content may be incorrect.

**Serverless SQL Pool**

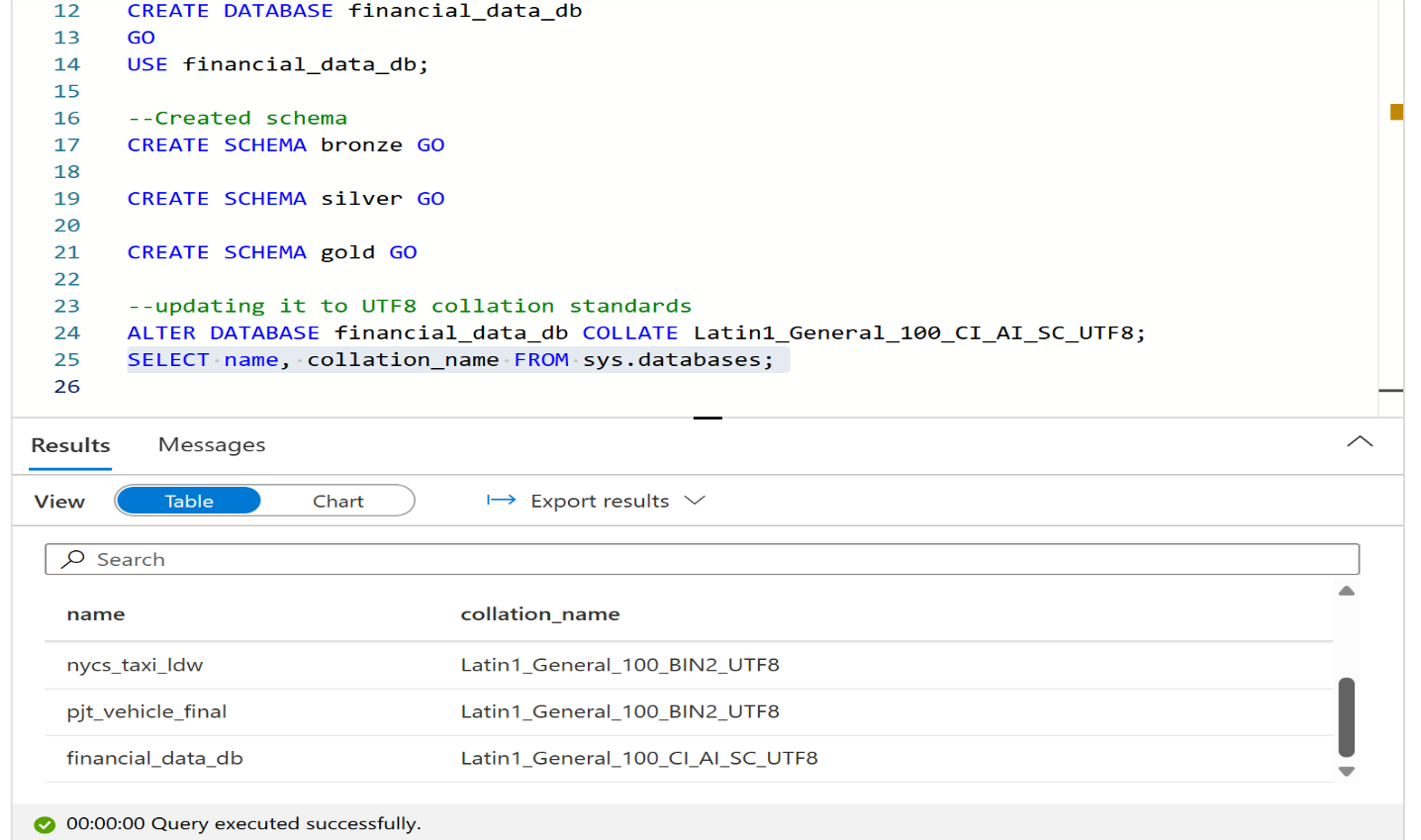
**A screenshot of a computer

AI-generated content may be incorrect.**

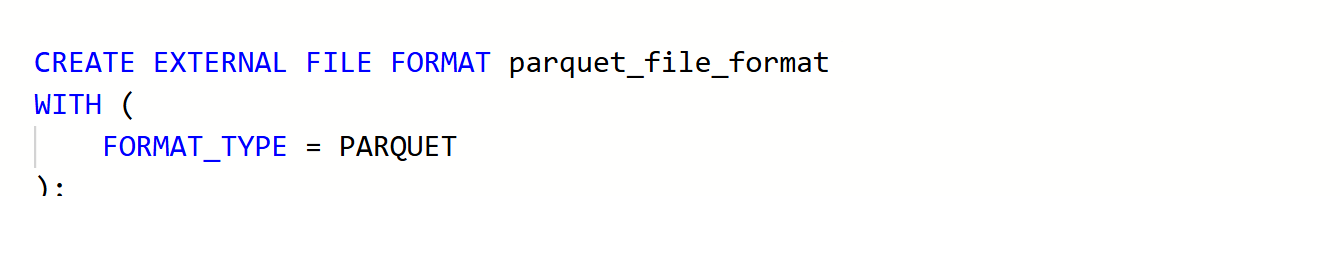
**Manually uploaded the csv file to the Synapse ADLS Gen 2**

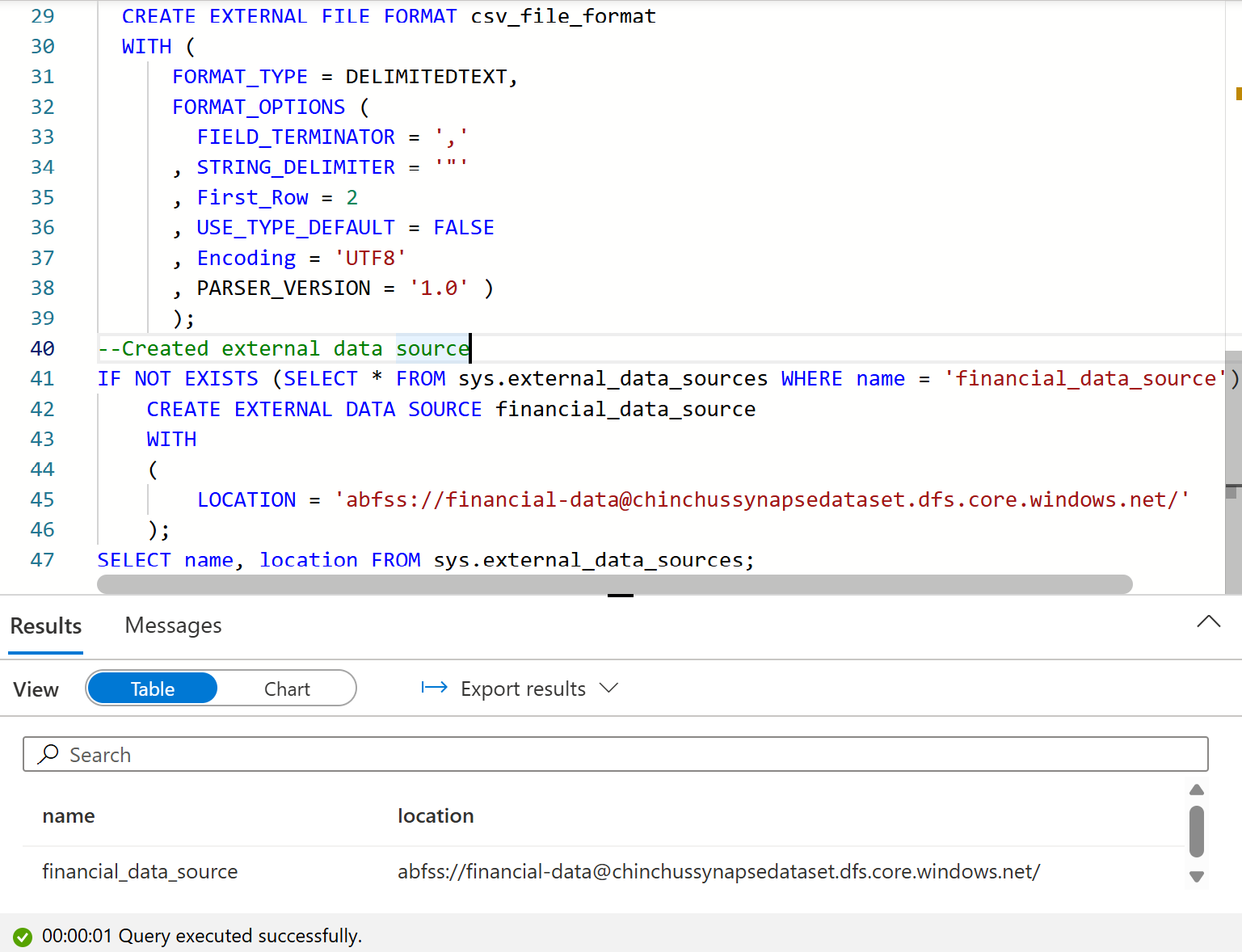
****

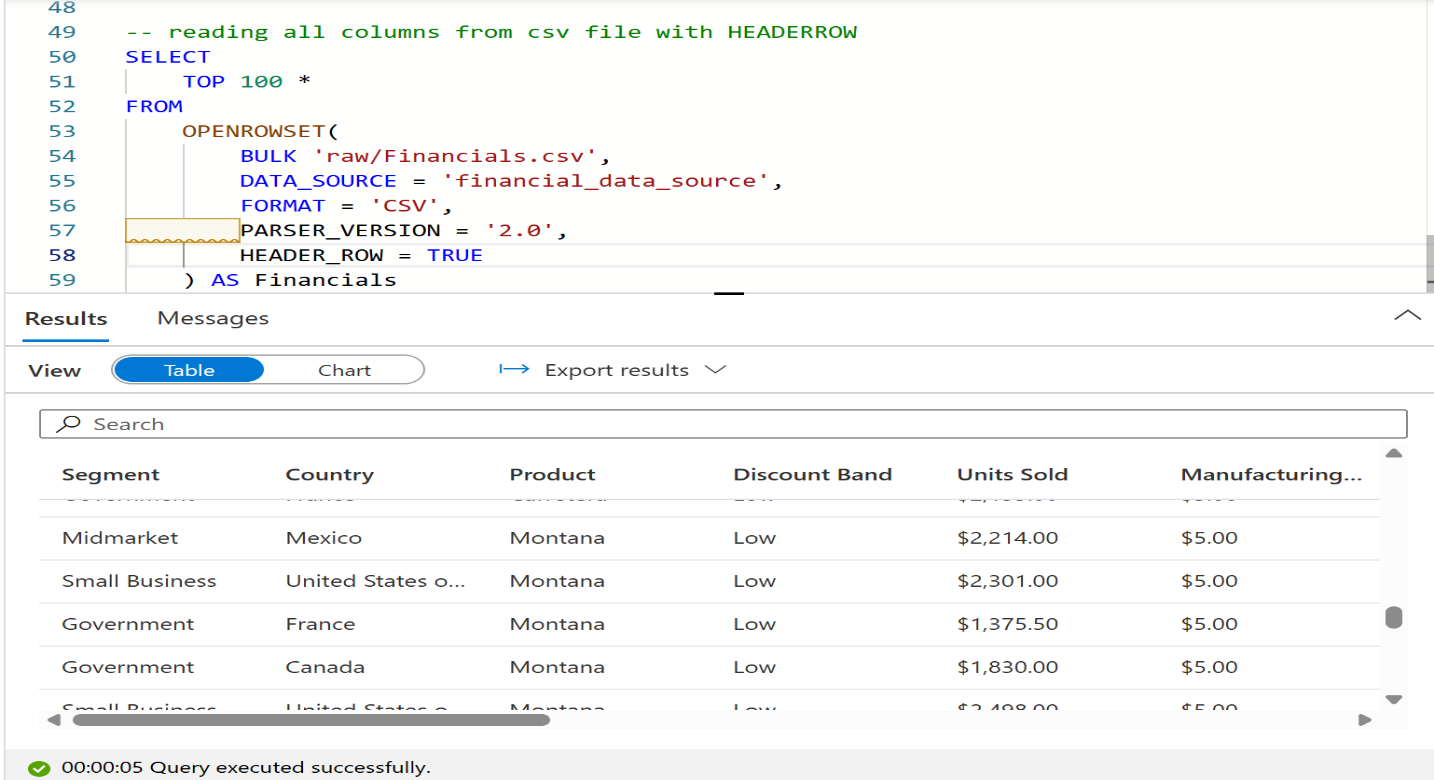
**Create Database and schema for 3 layers**

****

**Created external file format for both CSV and Parquet**

****

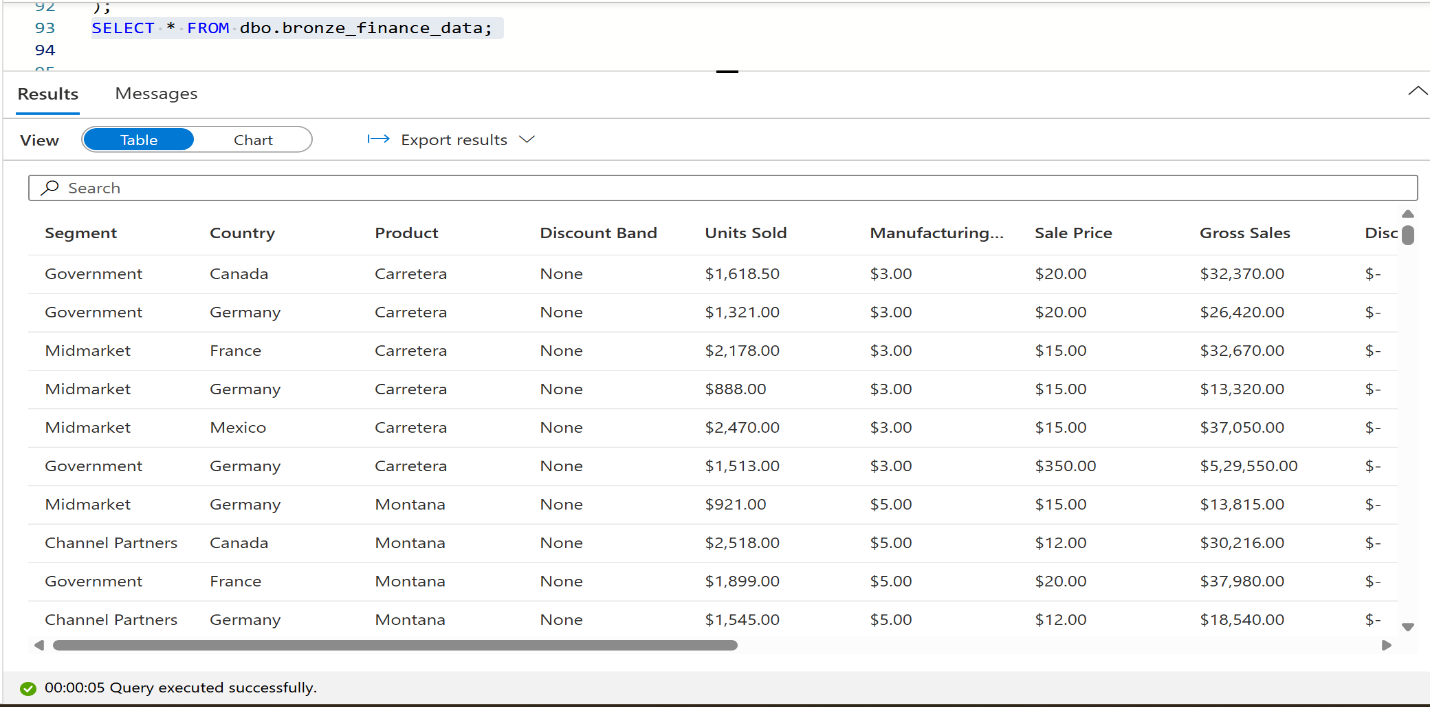
****

****

**Bronze Layer: External Table Creation**

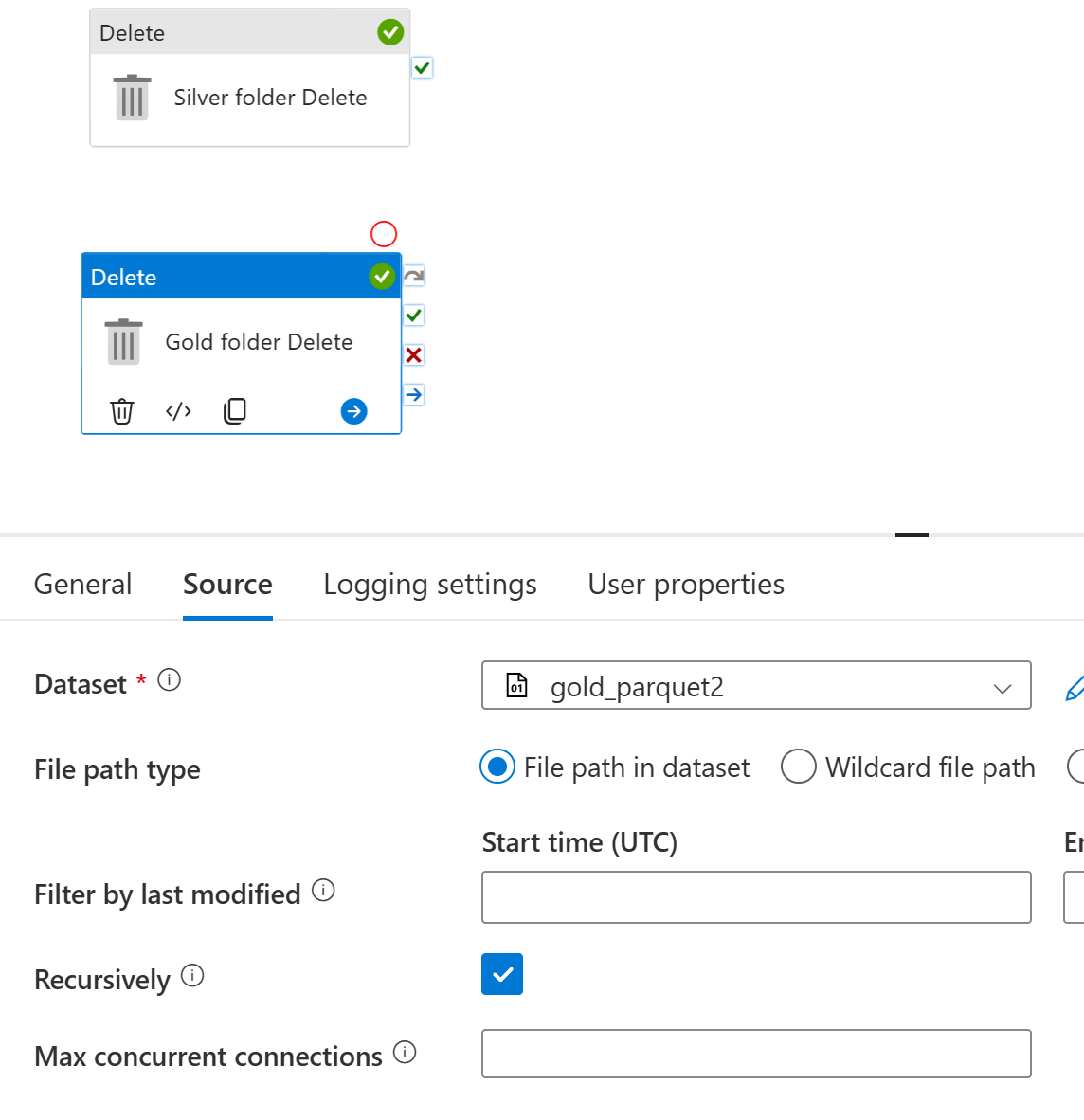
****

**Result**

****

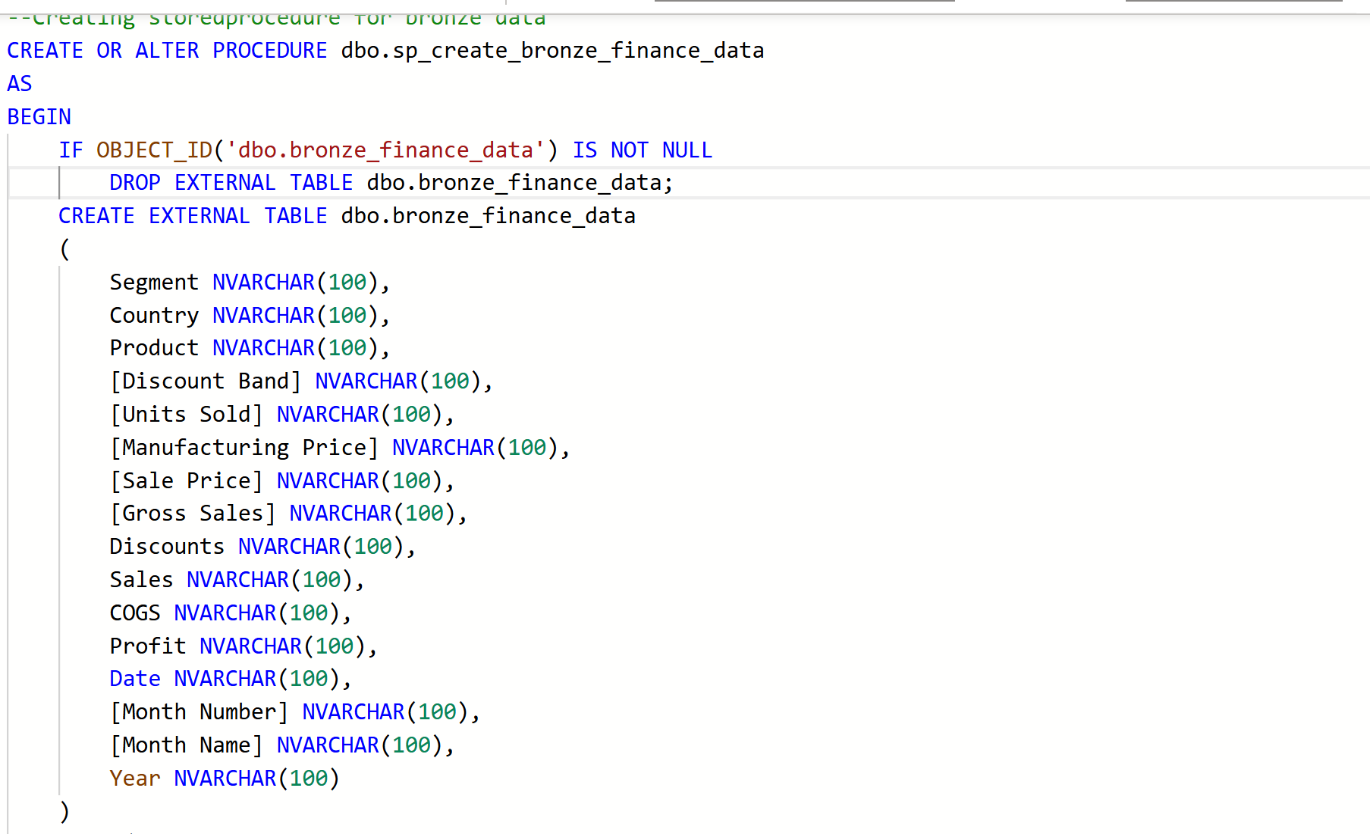
**Pipeline: Creating Delete activities for both silver and gold layer**

****

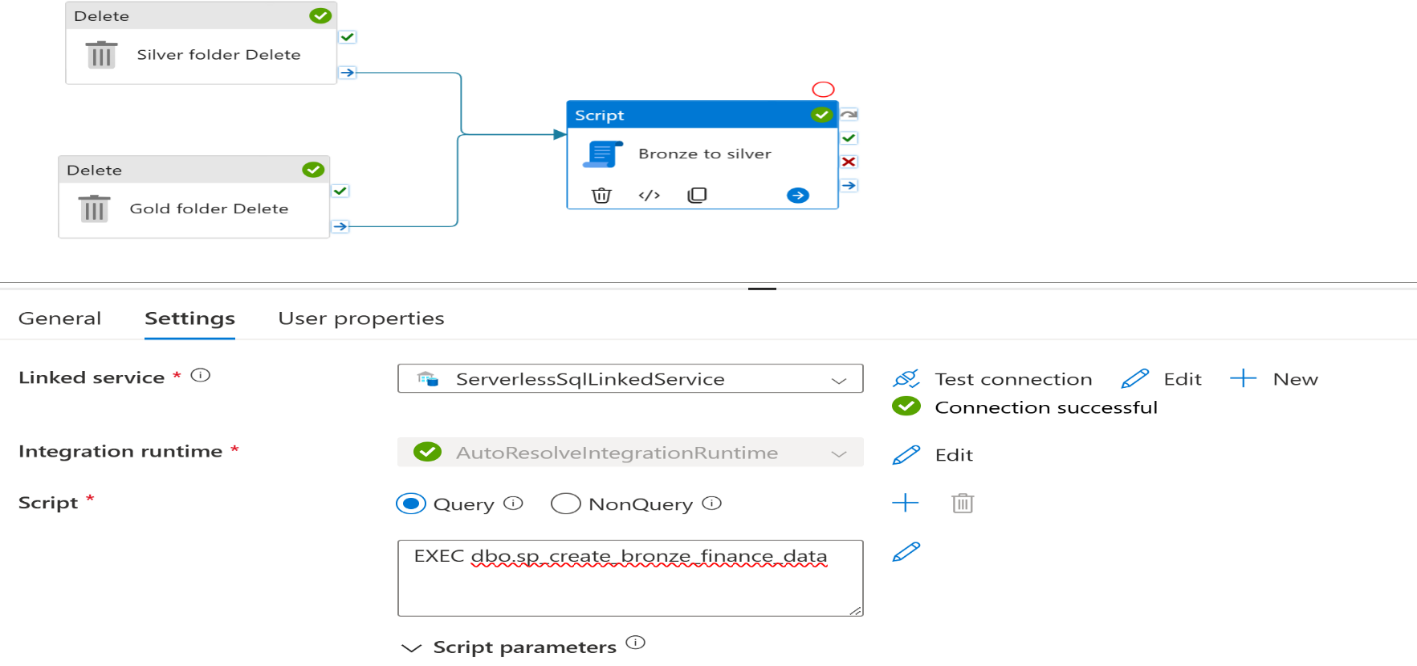
****

**Pipeline: Bronze Table Creation**

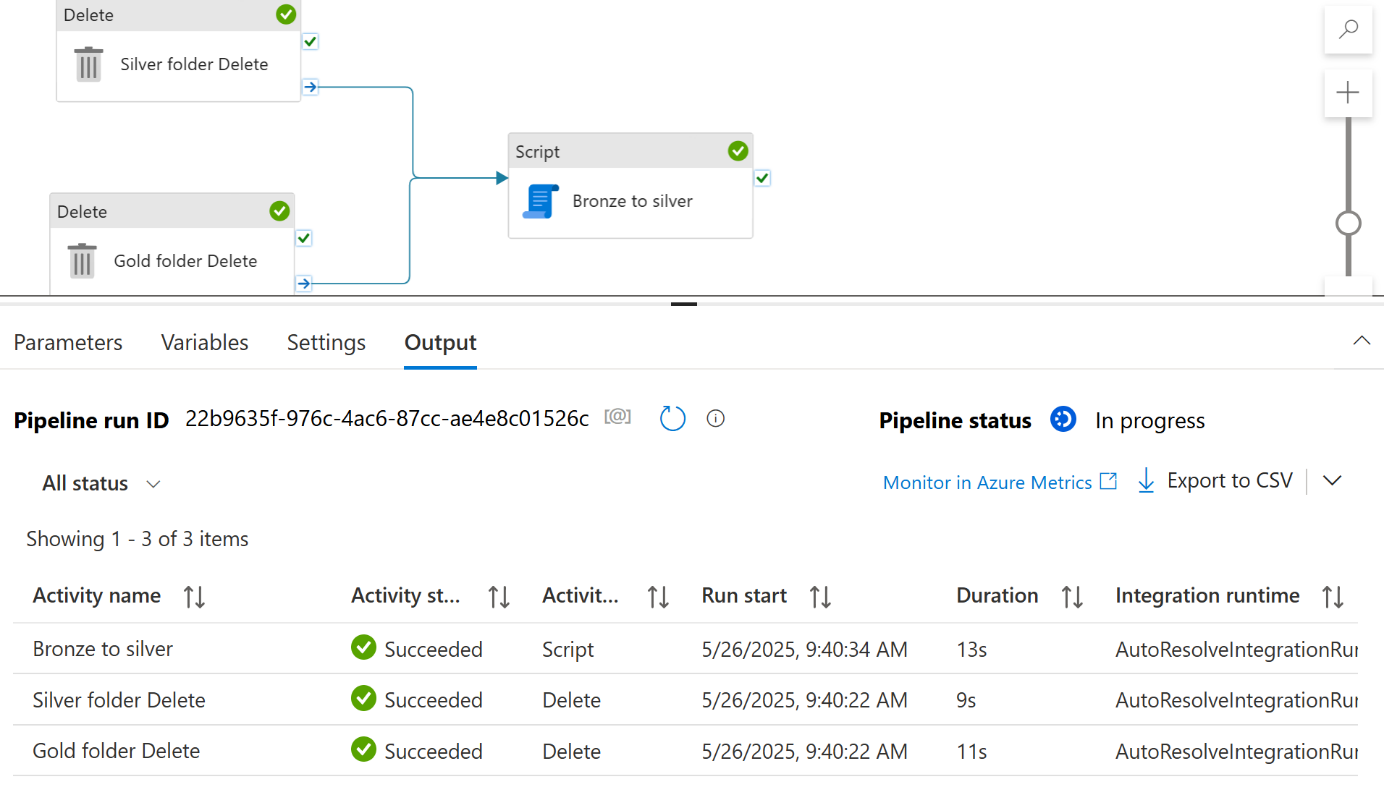
**Created a Stored Procedure for the bronze external table.**

****

**Added a new script Activity in the pipeline**

****

**Debug the pipeline**

****

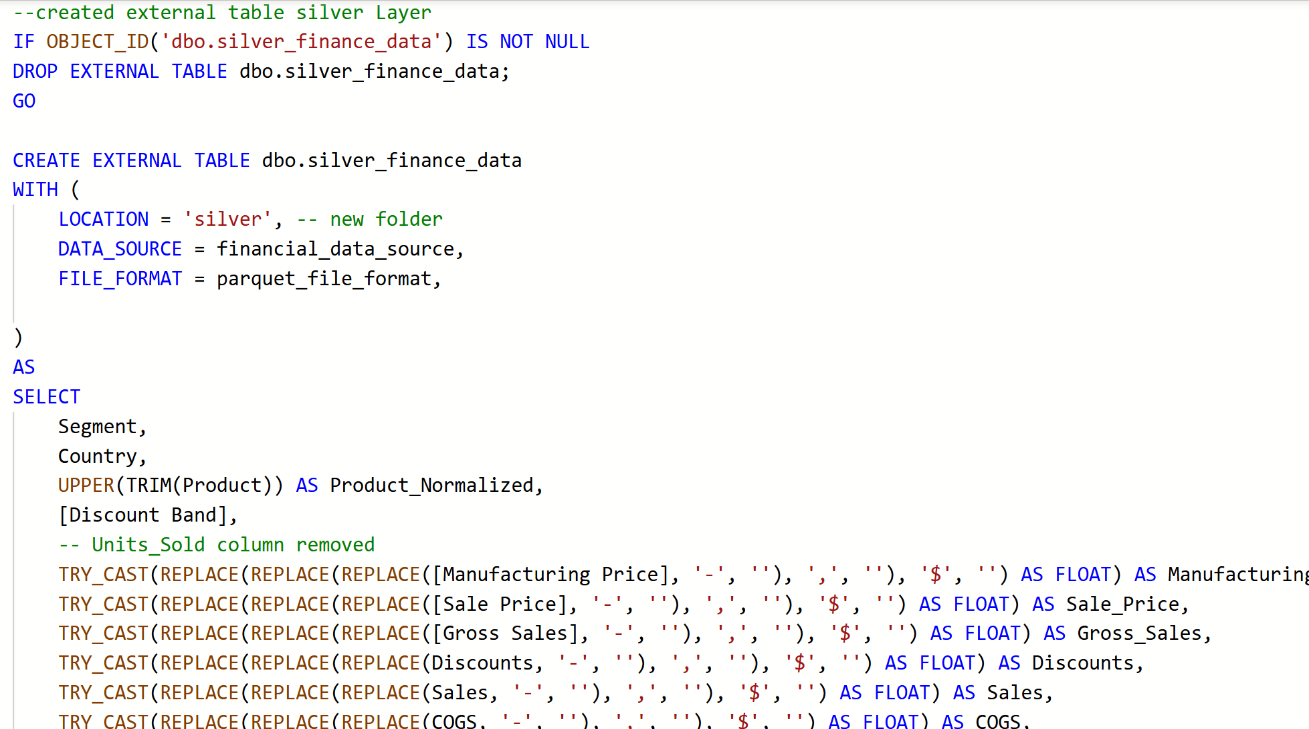
**Silver layer creation - Data Cleaning & basic transformations**

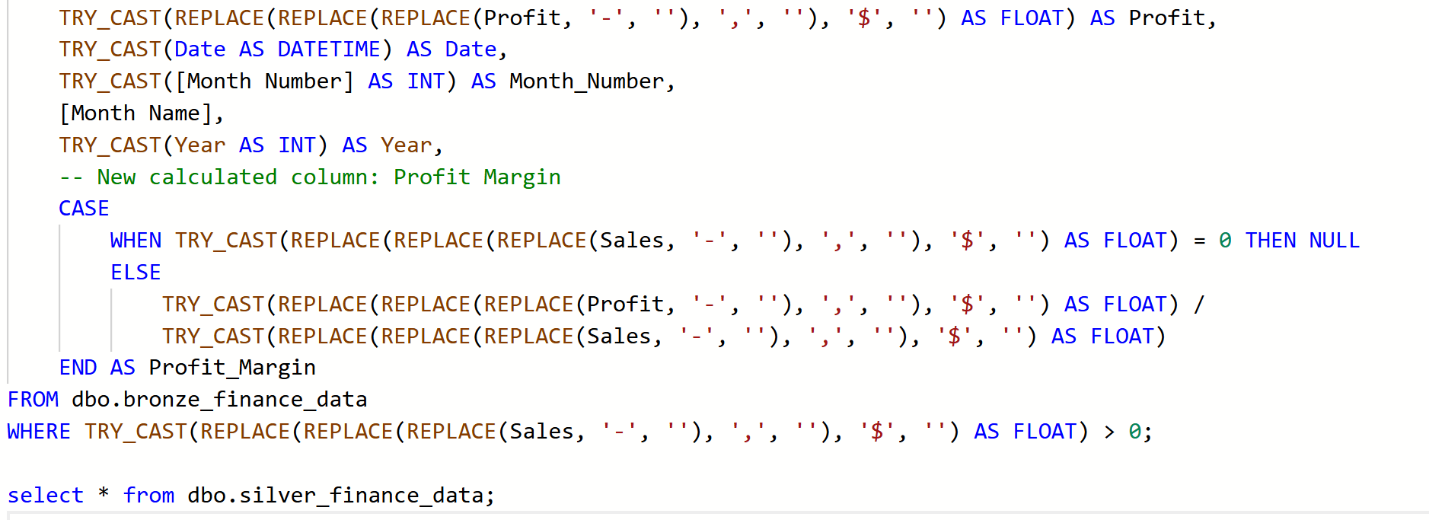
From the data viewed in MS Excel I have found that some basic cleaning needed which includes

* Remove '$' sign, “,” (comma) and '-' sign from all columns where they are present
* Change datatype from objects to int after the above two.
* A new column: **Profit Margin = Profit / Sales**
* Normalize product names (e.g., trim and make upper case)
* Remove rows with **Sales = 0** (to avoid divide-by-zero or irrelevant rows)

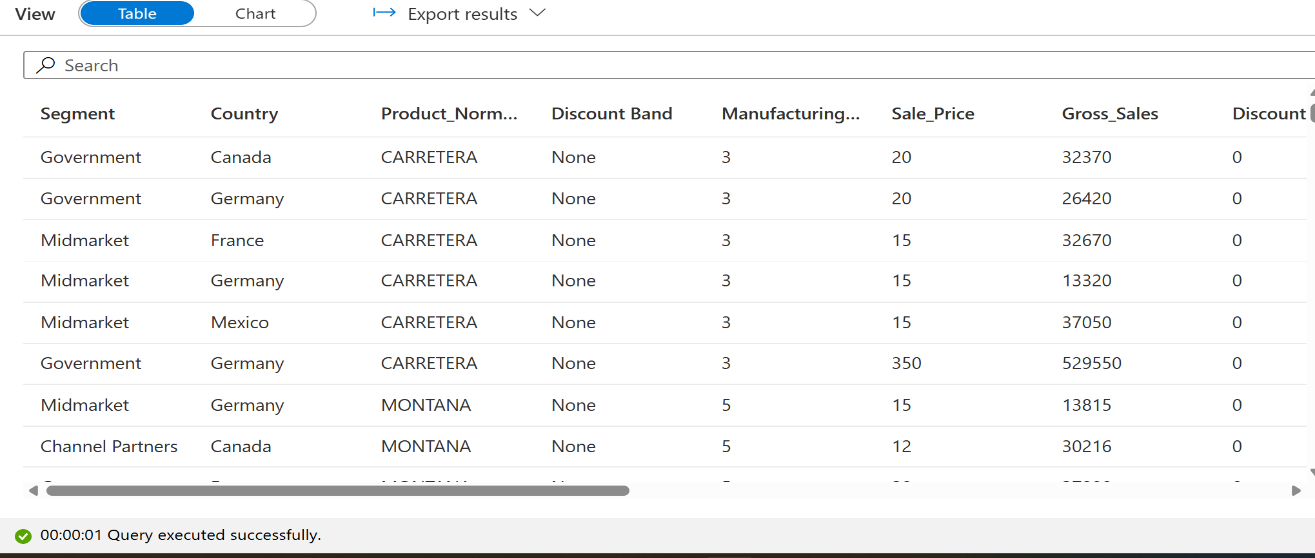
**Silver Table creation**

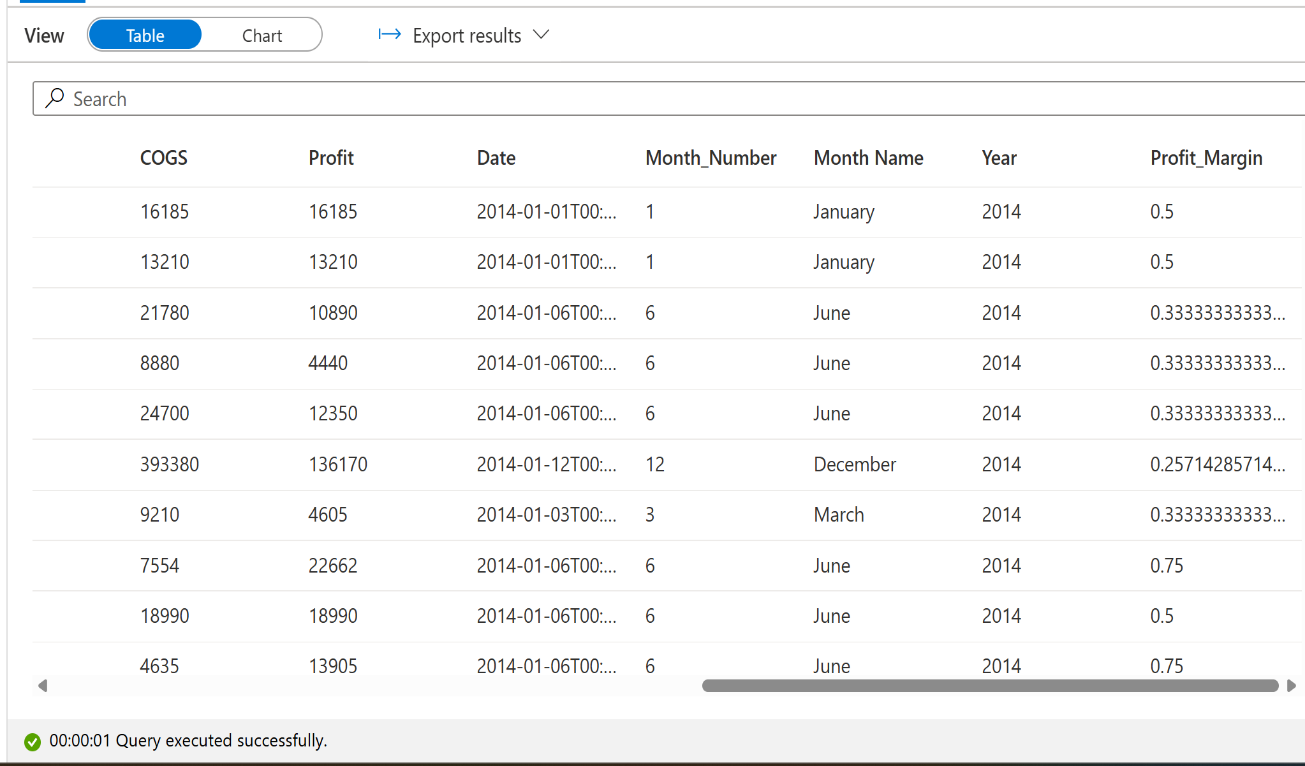
**Query**



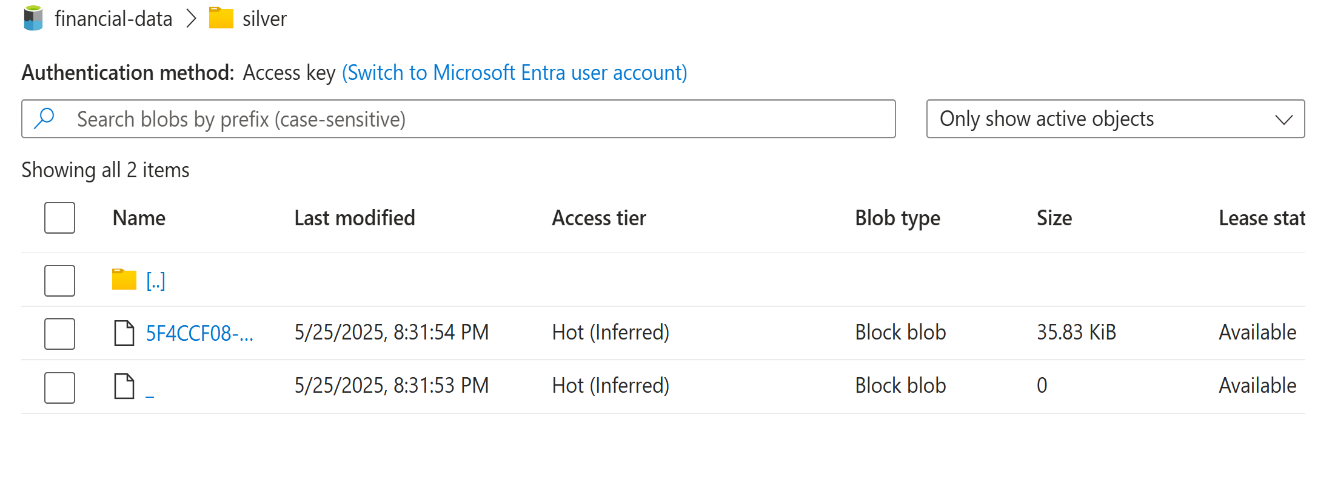


Result





**Now in my ADLS the cleaned and transformed data created in silver layer as parquet file.**



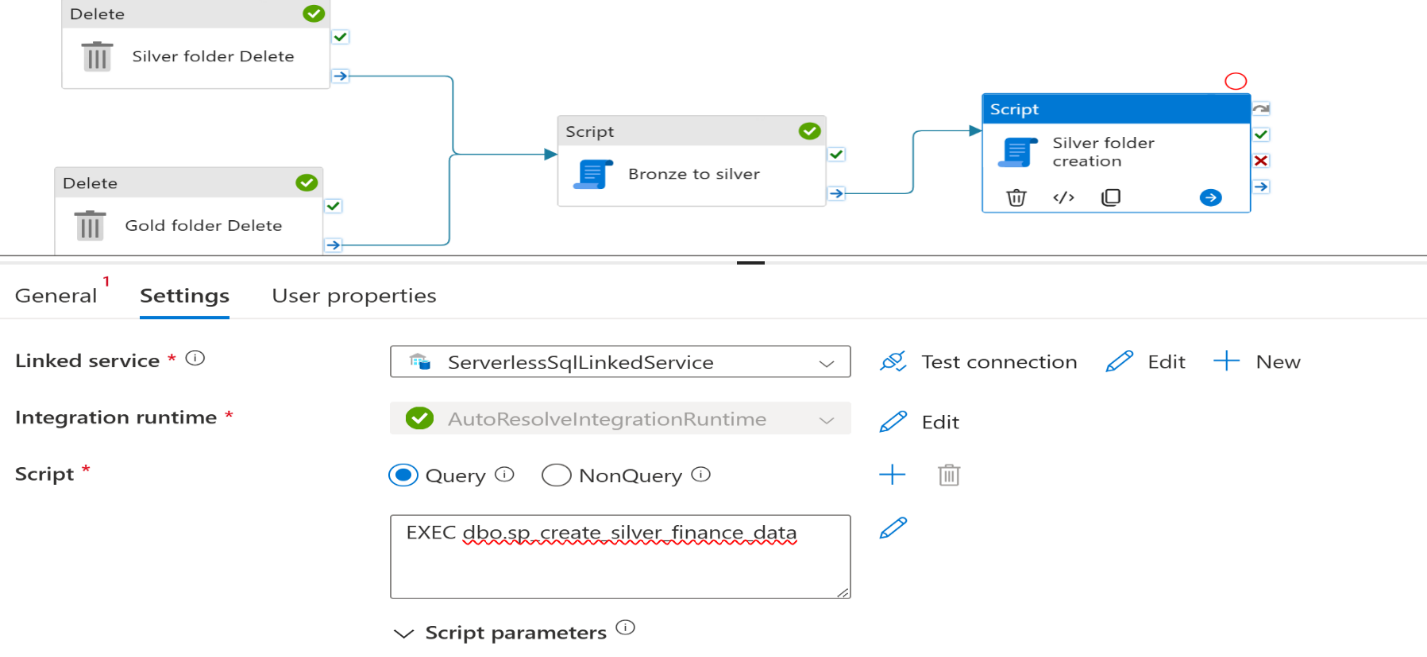
**Pipeline: Silver Folder Creation**

**Created another Stored Procedure for the silver external table.**

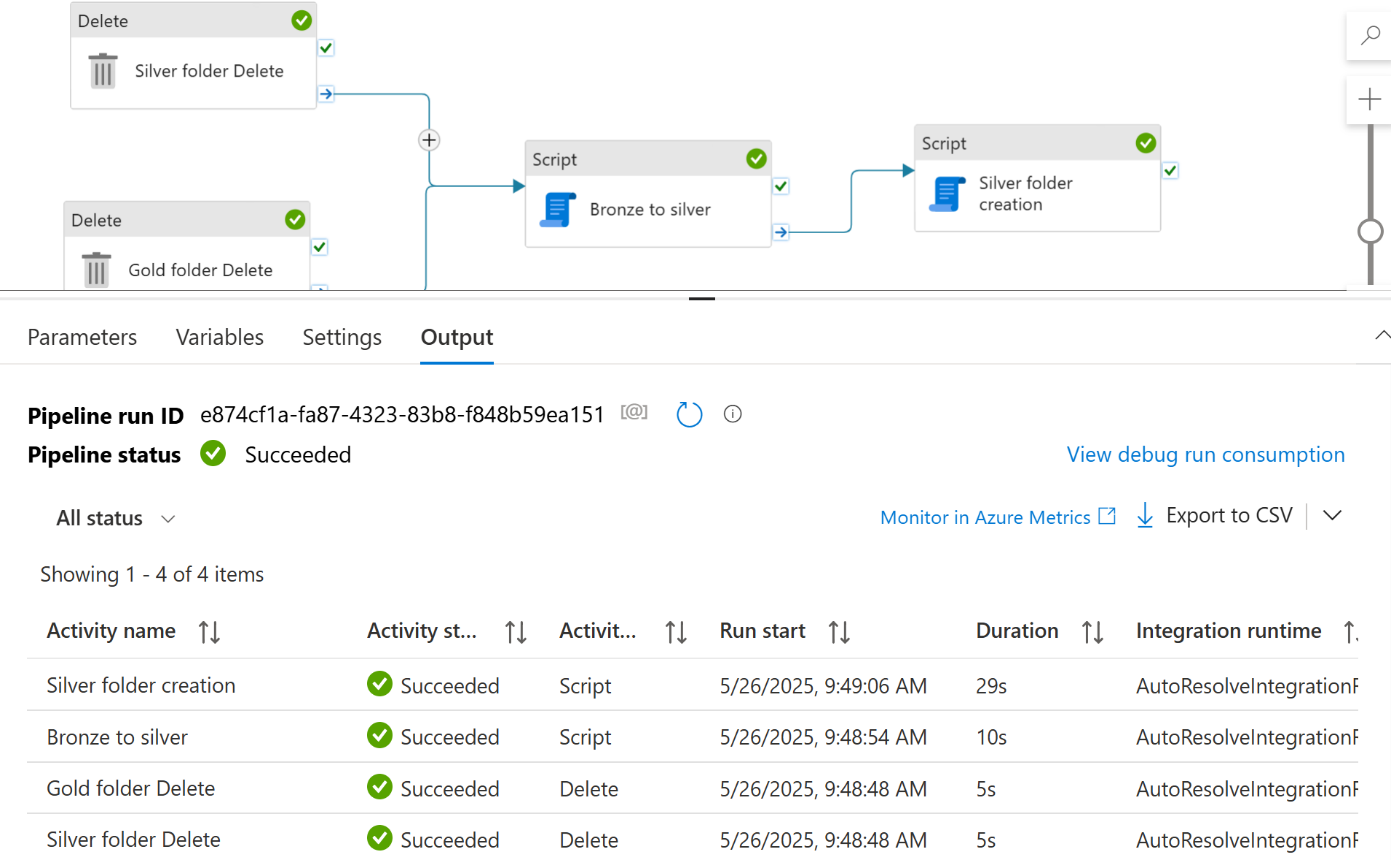
**A screen shot of a computer code

AI-generated content may be incorrect.**

**Added another script Activity in the pipeline**

****

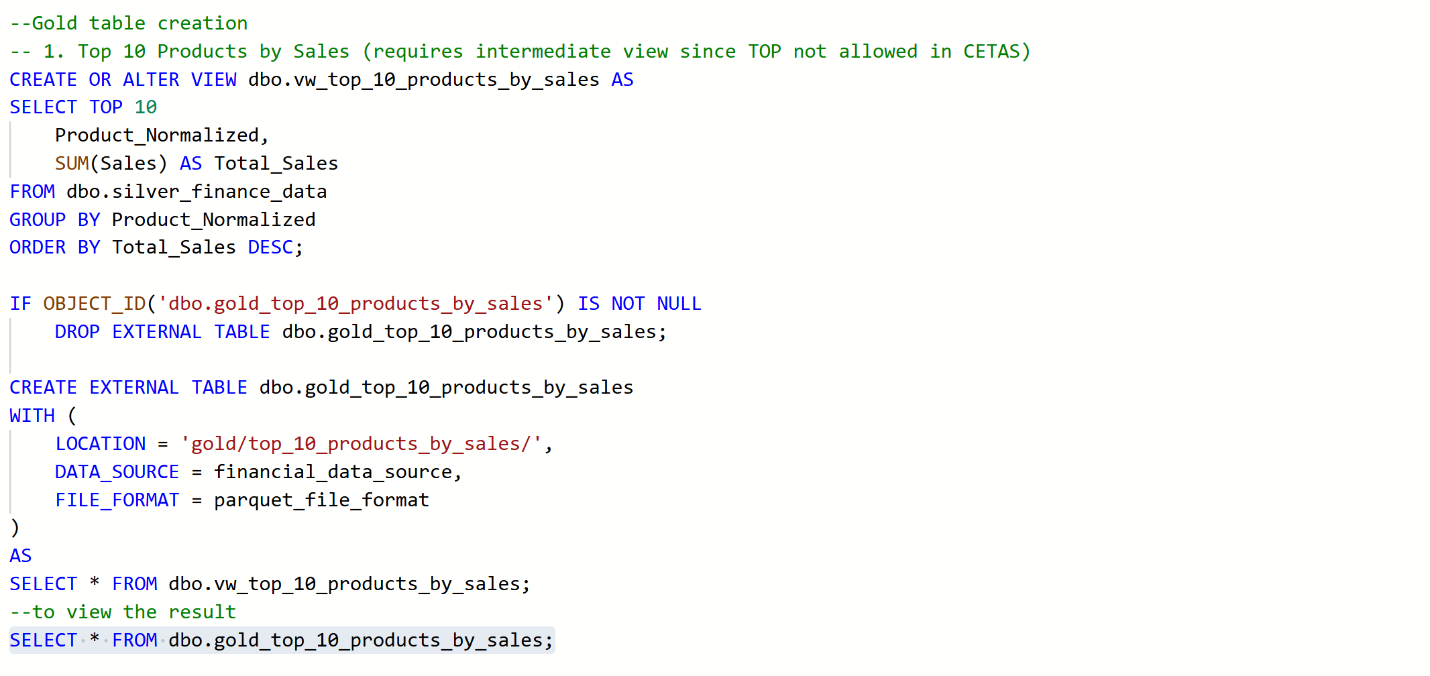
**Debug the pipeline**

****

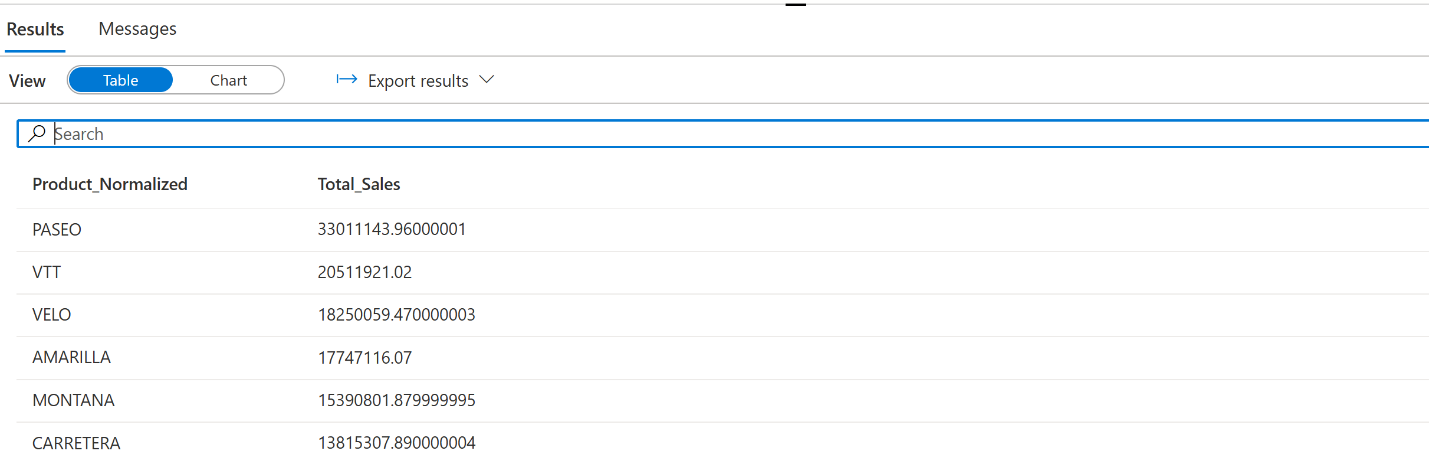
**Creating Gold layer Table**

**These incudes 3 tables**

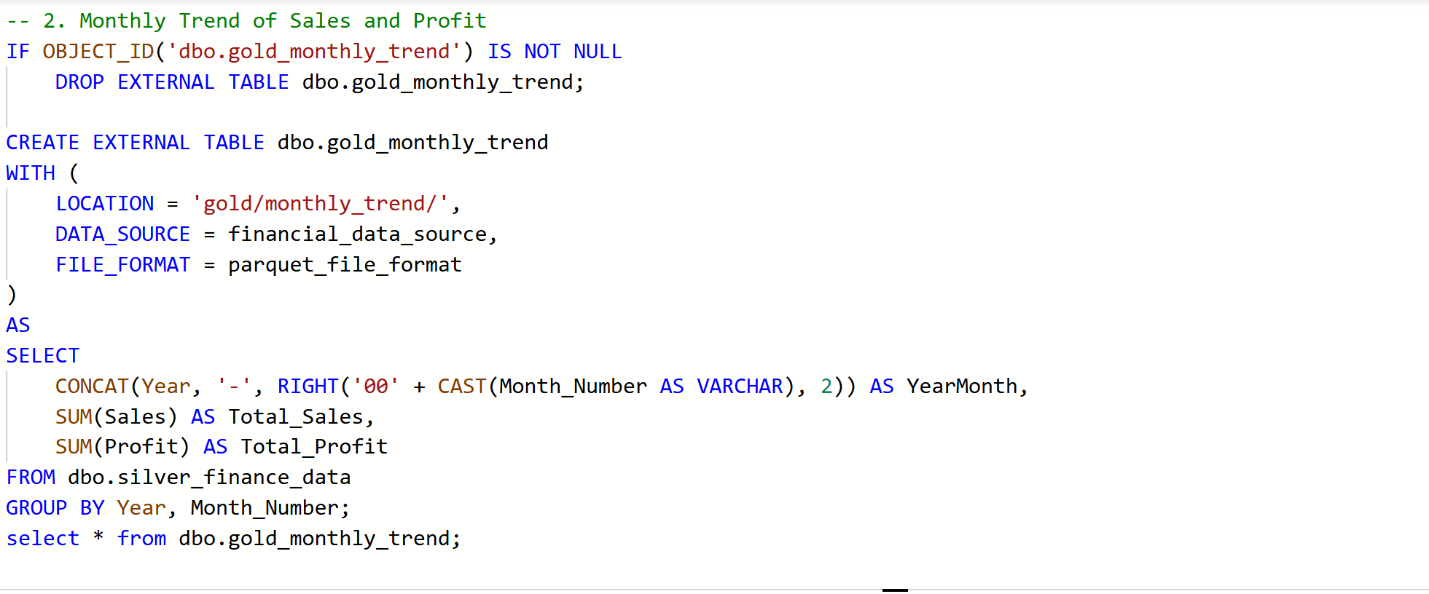
**Gold Table 1**

****

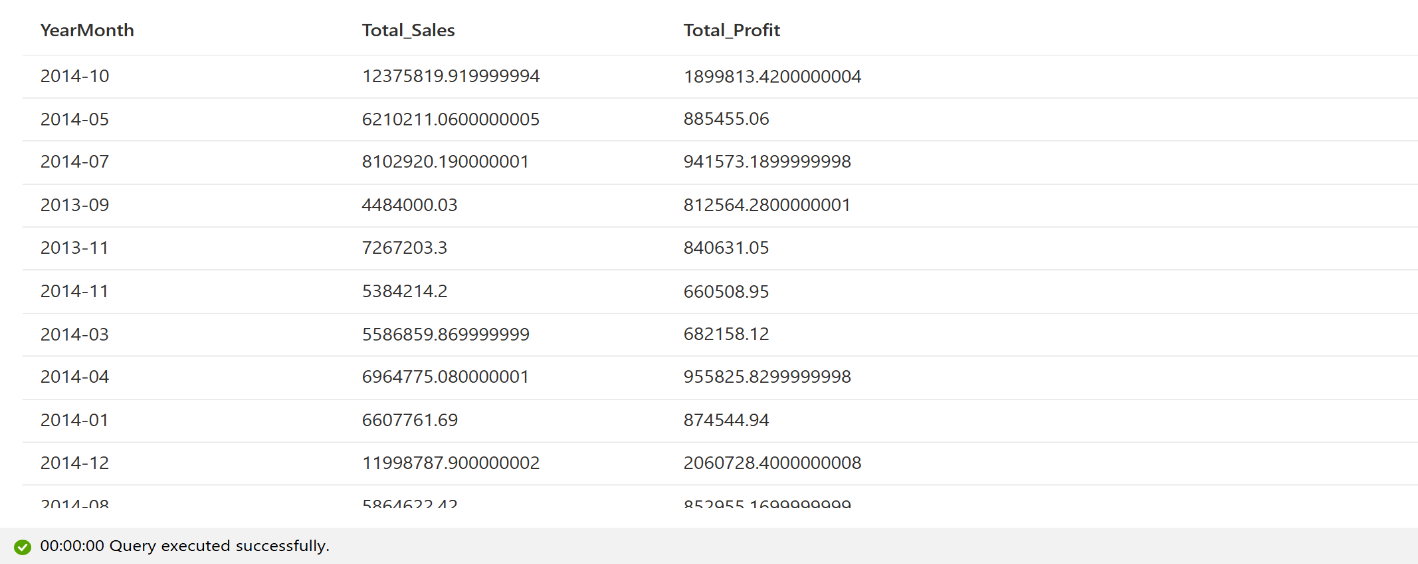
**Result:**

****

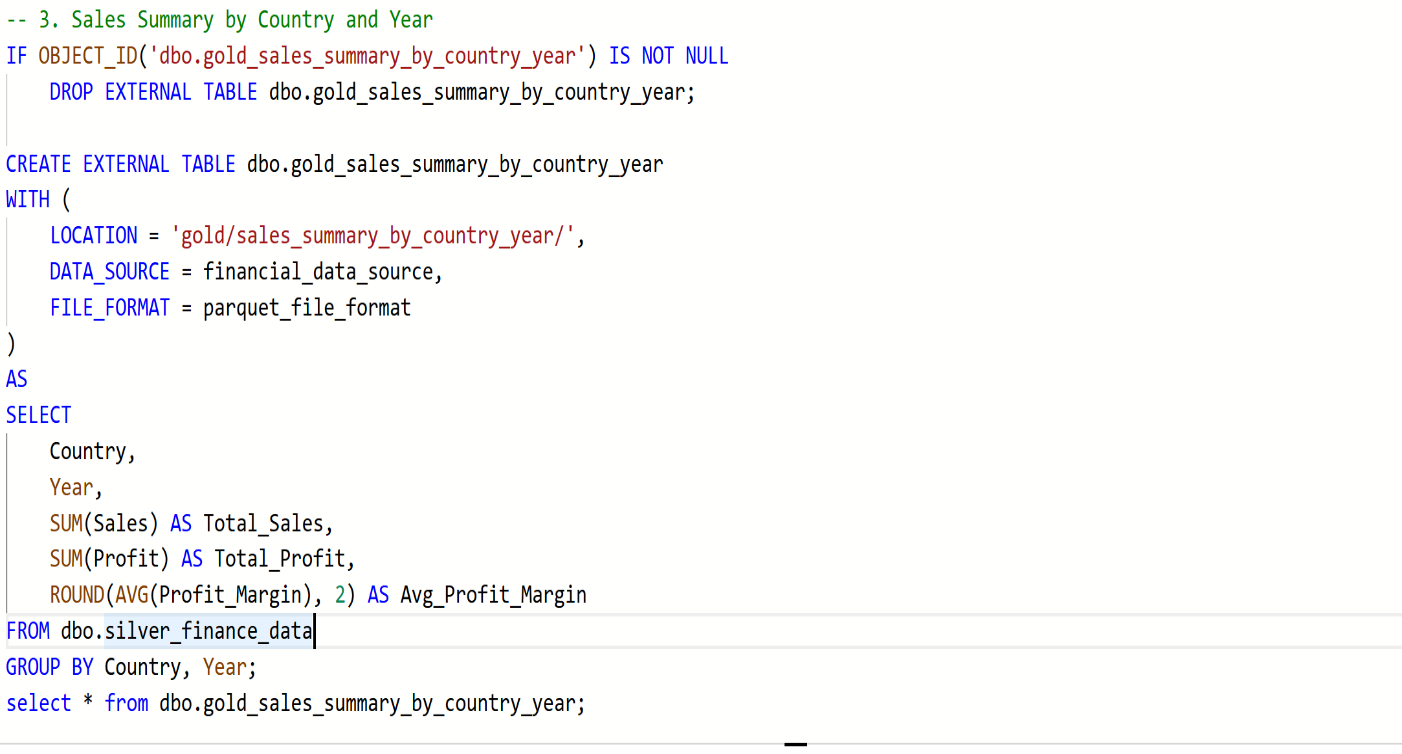
**Gold Table 2**

****

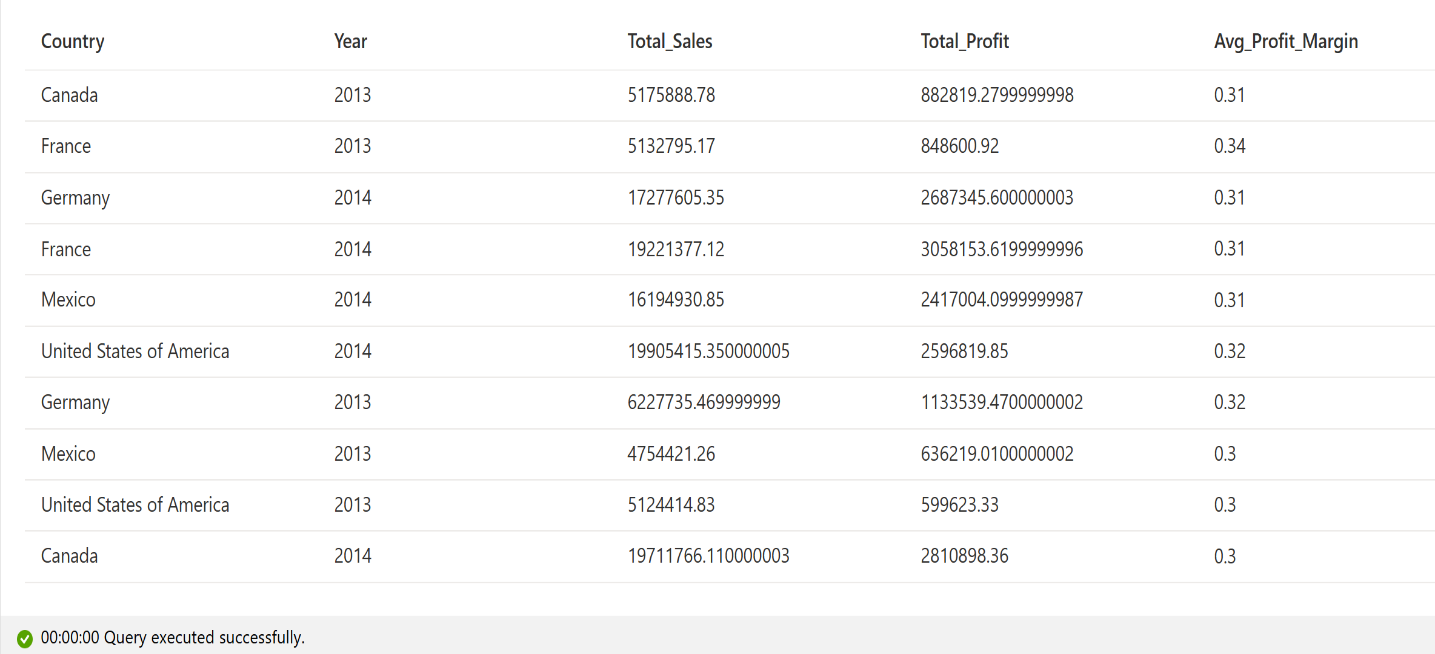
**Result:**

****

**Gold Table 3**

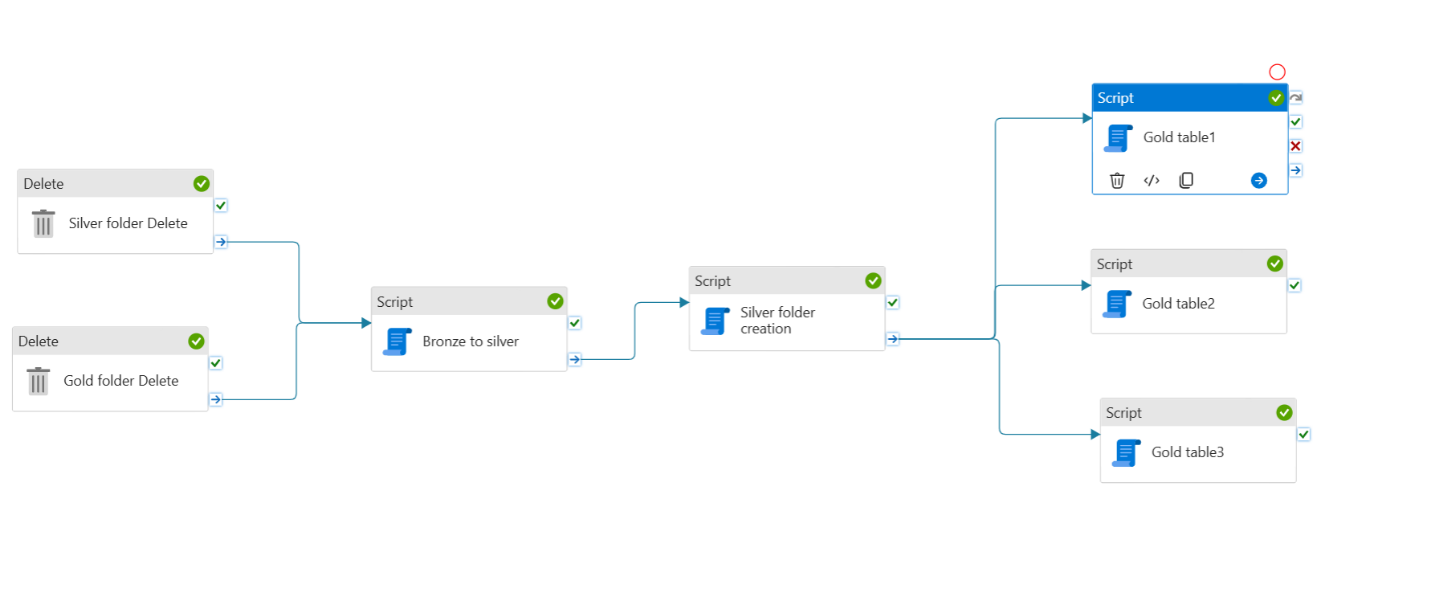
****

**Result**

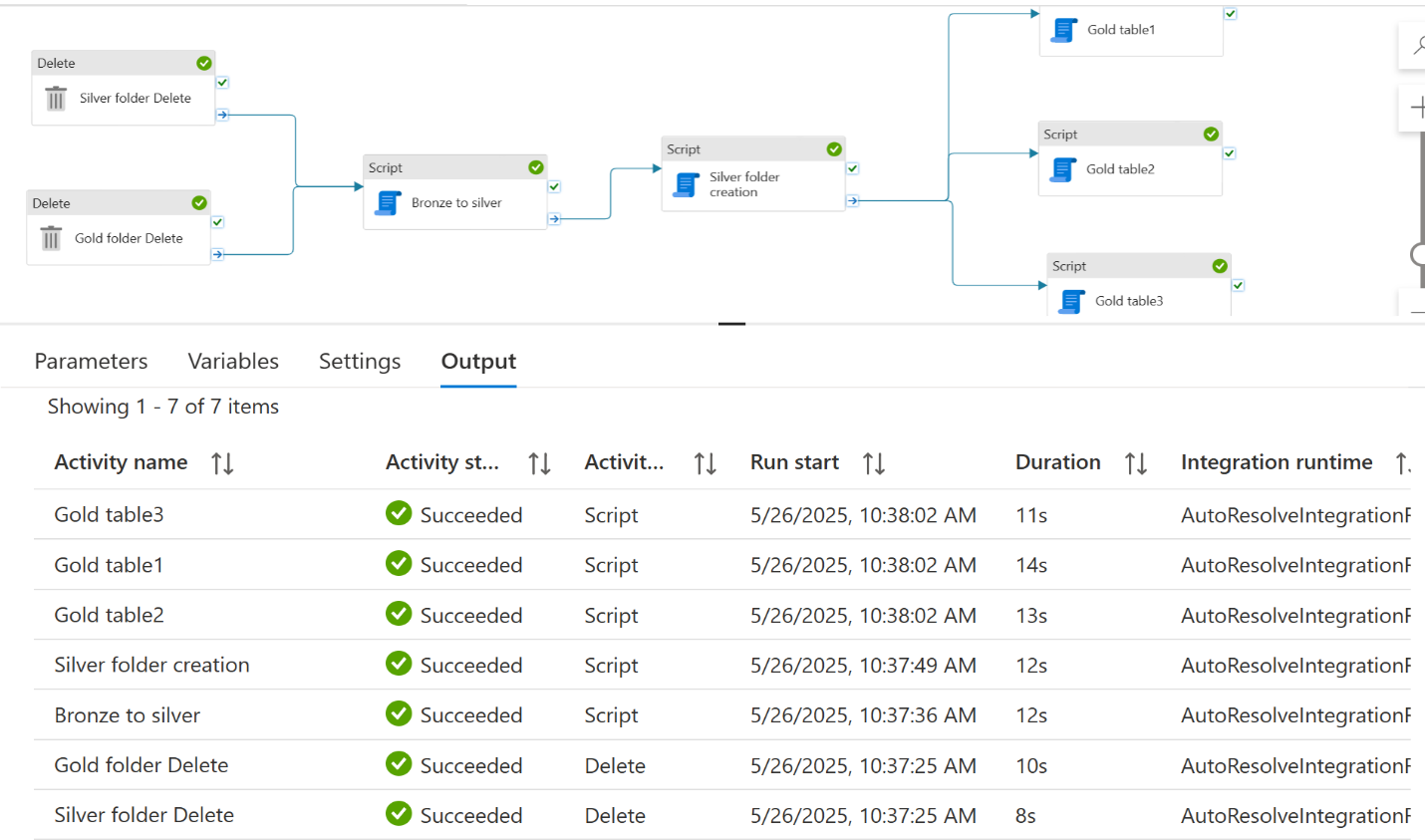
****

**Pipeline- Gold Layer**

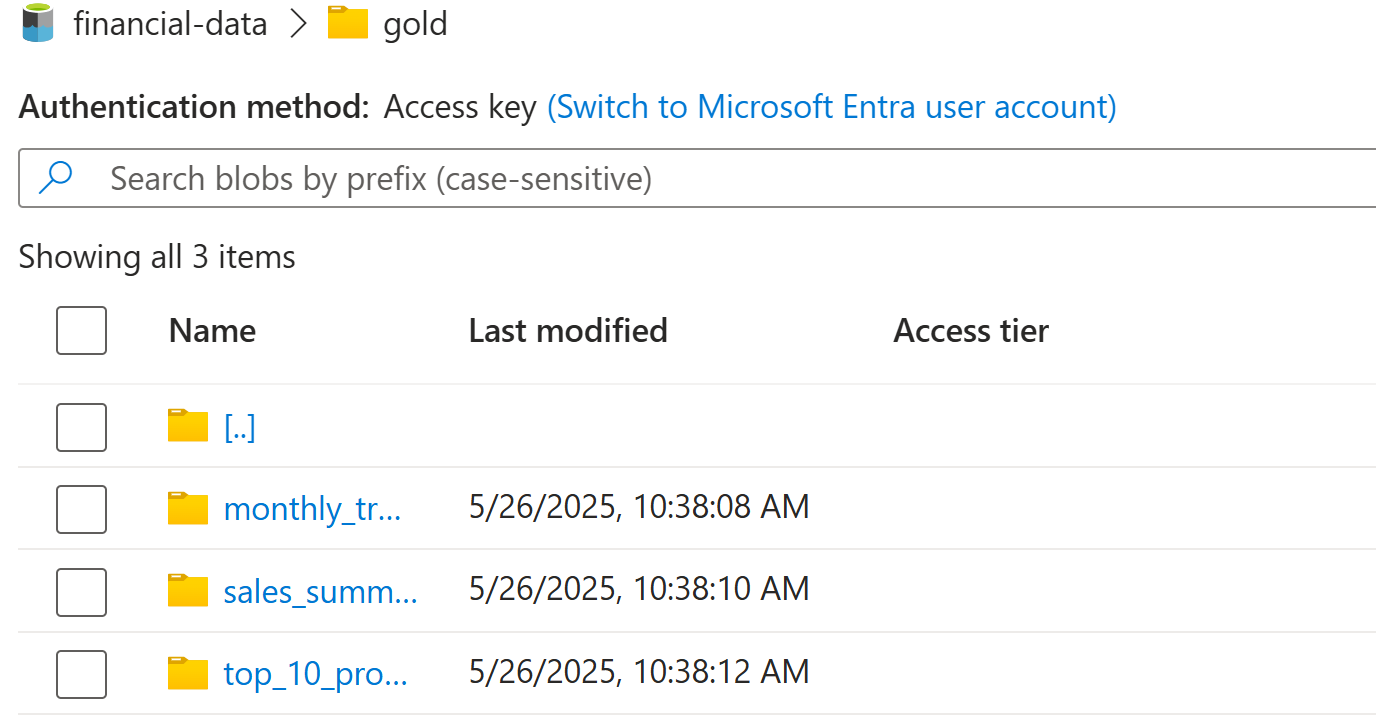
**Add 3 more script activities in this step for the above 3 tables in the gold layer.**

****

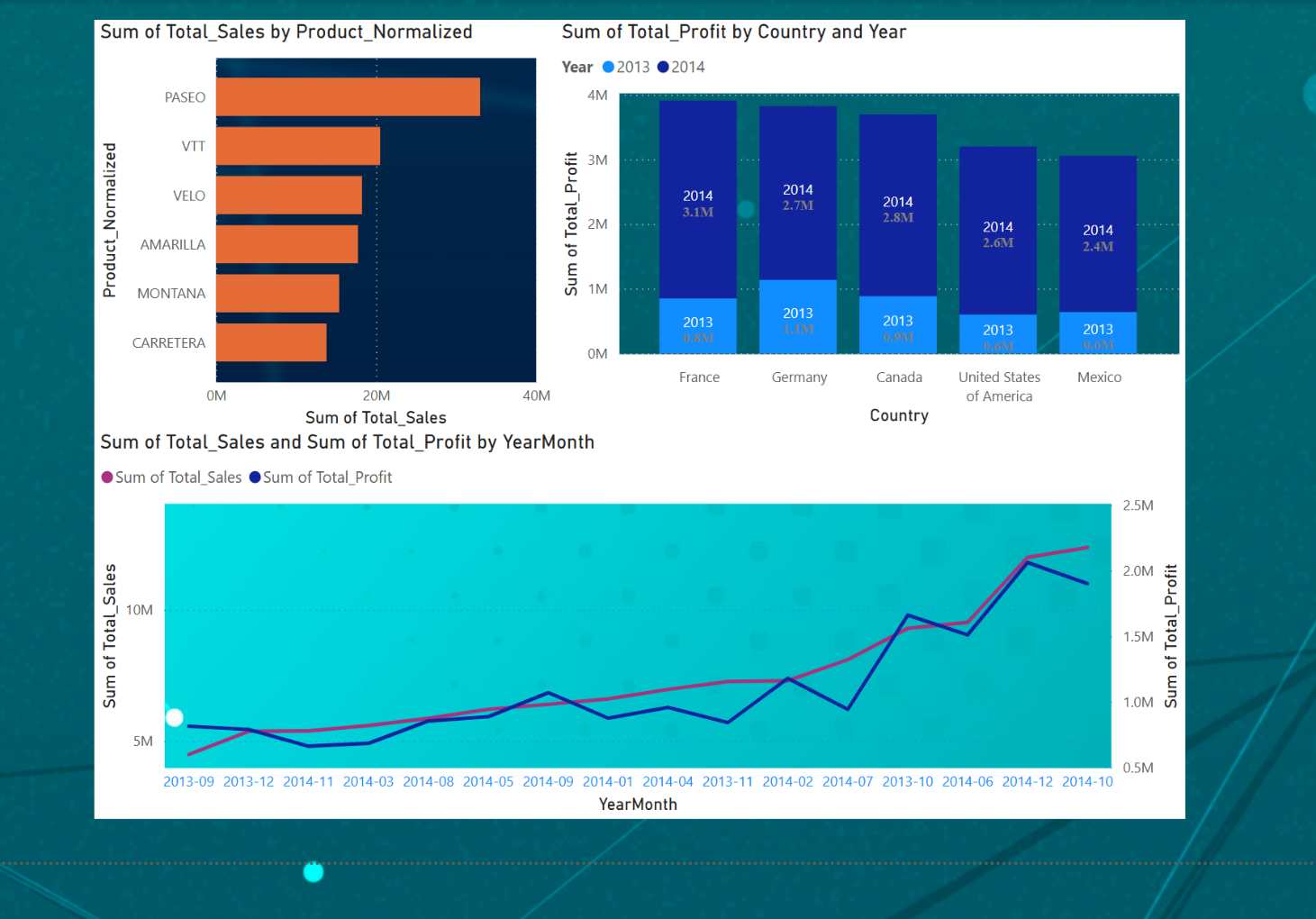
**Debug the pipeline**

****

**Now my ADLS container looks like**

****

**Company Financial sales Report (By profit)**

****

**Based on the various Tables created in the Gold layer of the process I have found that:**

**1. gold\_top\_10\_products\_by\_sales**

Visual: Horizontal bar chart showing Product\_Normalized vs Total\_Sales

Insights:

**Top-Selling Products**

* **PASEO** leads significantly in total sales, followed by **VTT** and **VELO**.
* Indicates strong product performance and demand concentration in a few key items.

**2**. **gold\_sales\_summary\_by\_country\_year**

Visual: Stacked column chart of Total\_Profit by Country split across Year

Insights:

**Profit by Country and Year**

* **France** had the highest profit in both 2013 and 2014, growing from **0.8M to 3.1M**.
* All countries saw a significant increase in 2014 compared to 2013, likely due to market expansion or product strategy shifts.
* **Mexico** showed relatively lower profit margins, suggesting potential areas for strategic improvement.

**3. gold\_monthly\_trend**

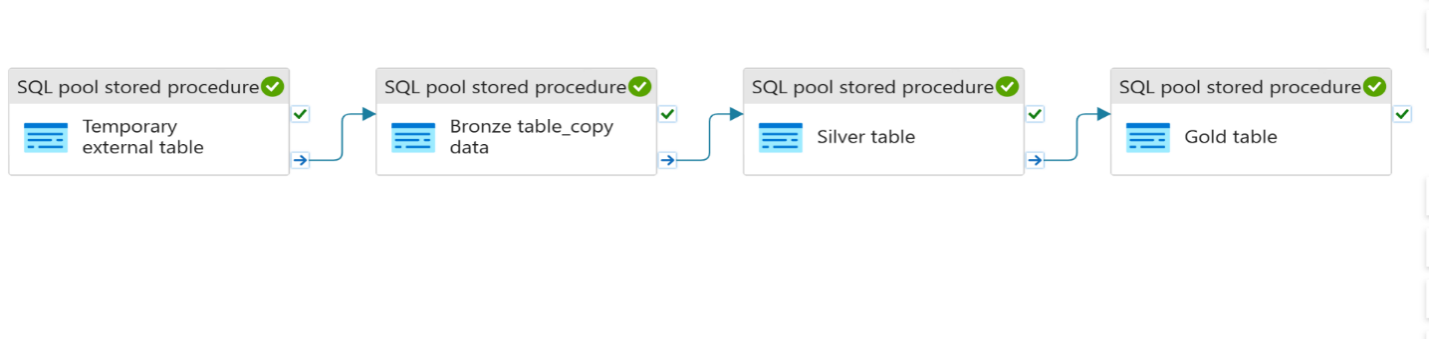
Visual: Dual-line graph for Total\_Sales and Total\_Profit over YearMonth

Insights:

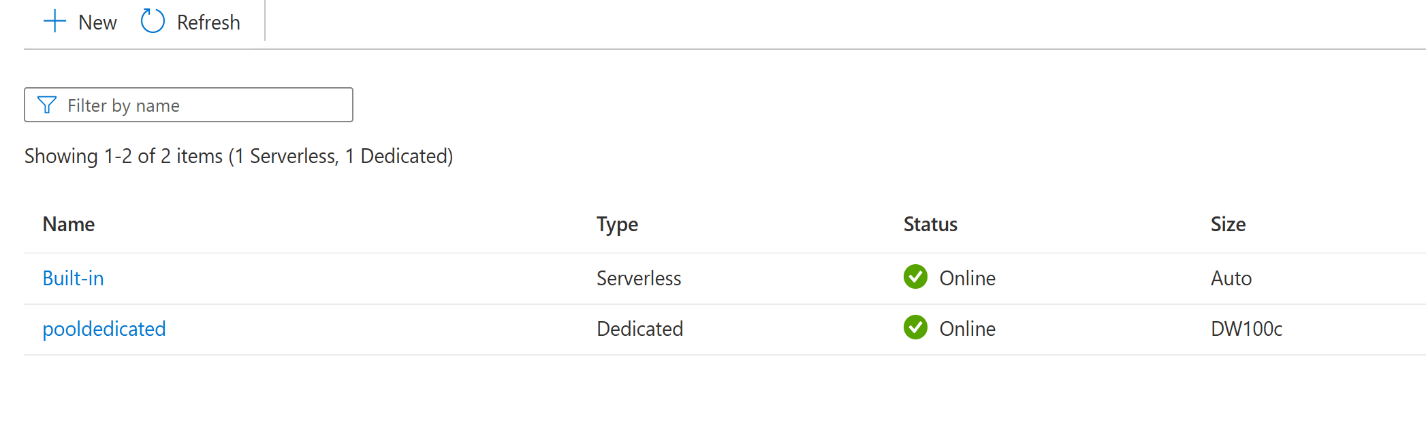
**Time Series Trends (Sales & Profit)**

* Gradual and consistent increase in both **Total Sales** and **Total Profit** from **2013-09 to 2014-10**, Peak activity observed in **2014-09 and 2014-10**, potentially due to seasonal campaigns or holiday demand.
* Sales and profit generally move in tandem, indicating efficient cost management.

**Dedicated SQL Pool**

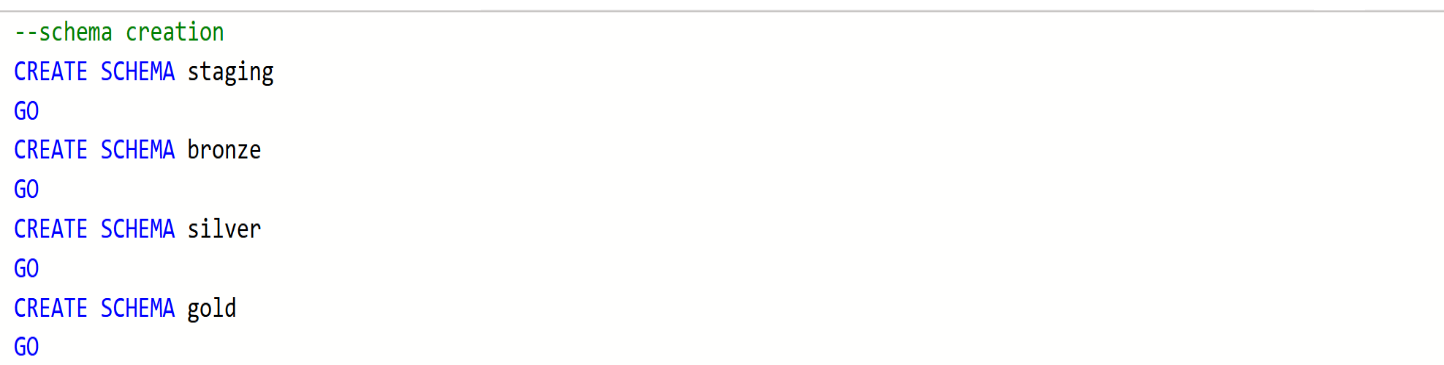
****

**Created new dedicated sql pool**

**Bronze Layer setup**

As dedicated SQL pool acts as the database, not need to create database

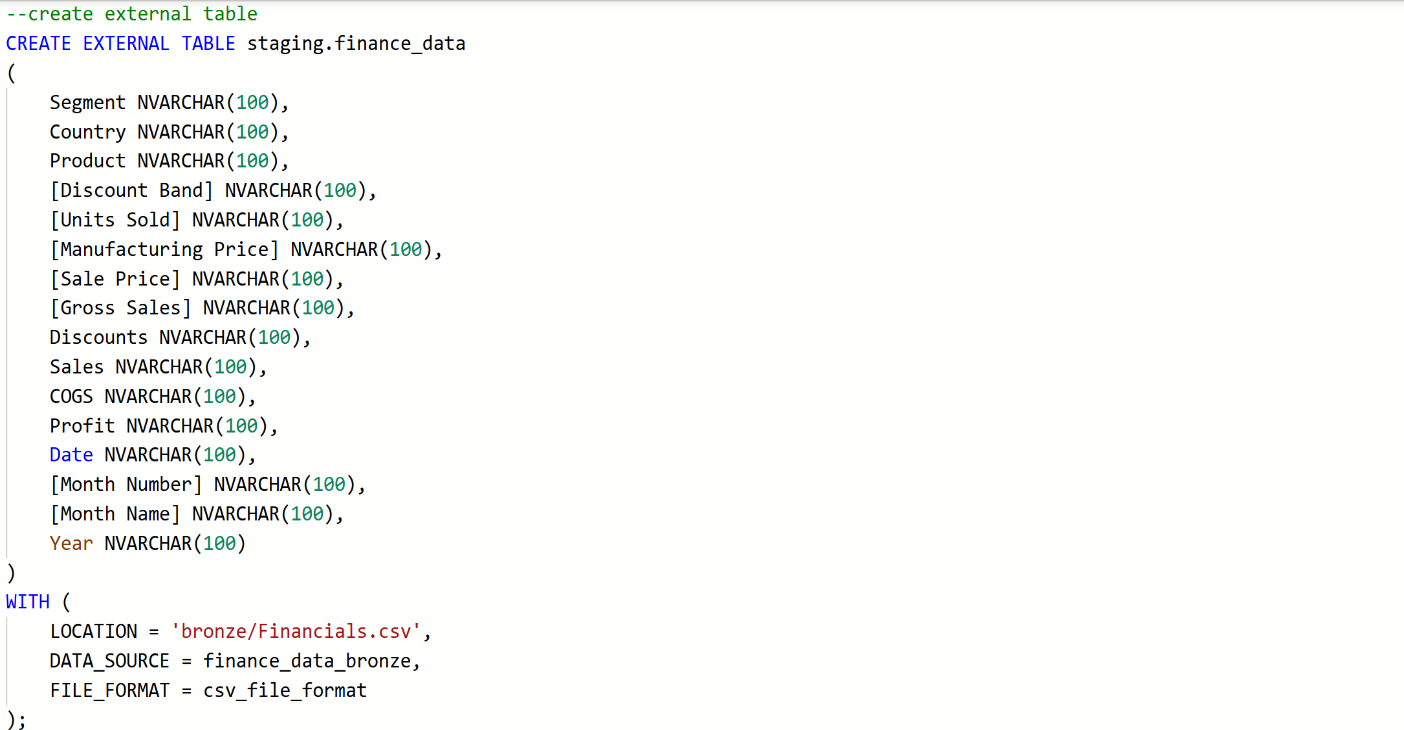
1 Created schemas for all layers



**Created external data source and file format**



**Created a temporary external table (for staging)**



****

**Result**:



**Automation using Pipeline (Dedicated\_finance\_data\_pipeline)**

**Temporary external table creation**

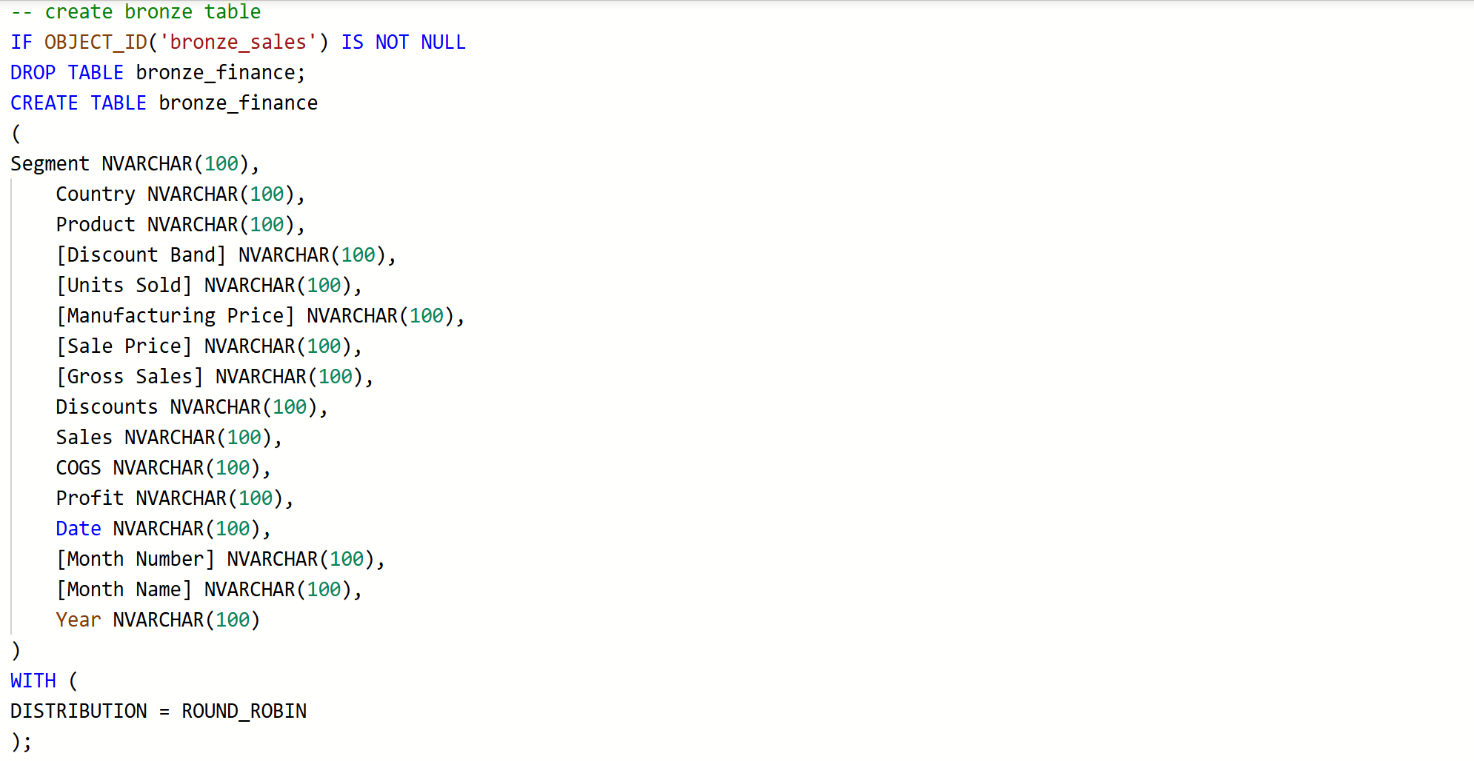


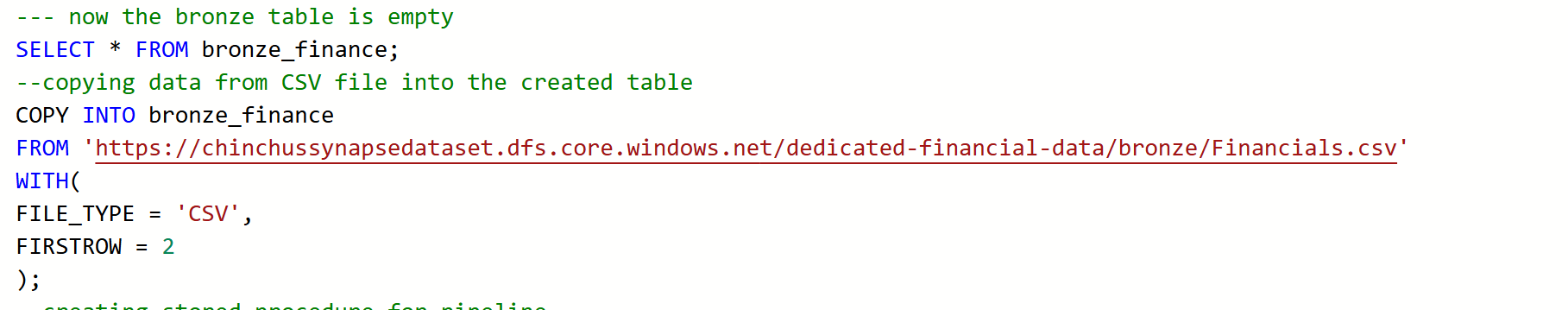
**Added a Sql pool stored procedure activity and did the debugging process**

A screenshot of a computer

AI-generated content may be incorrect.

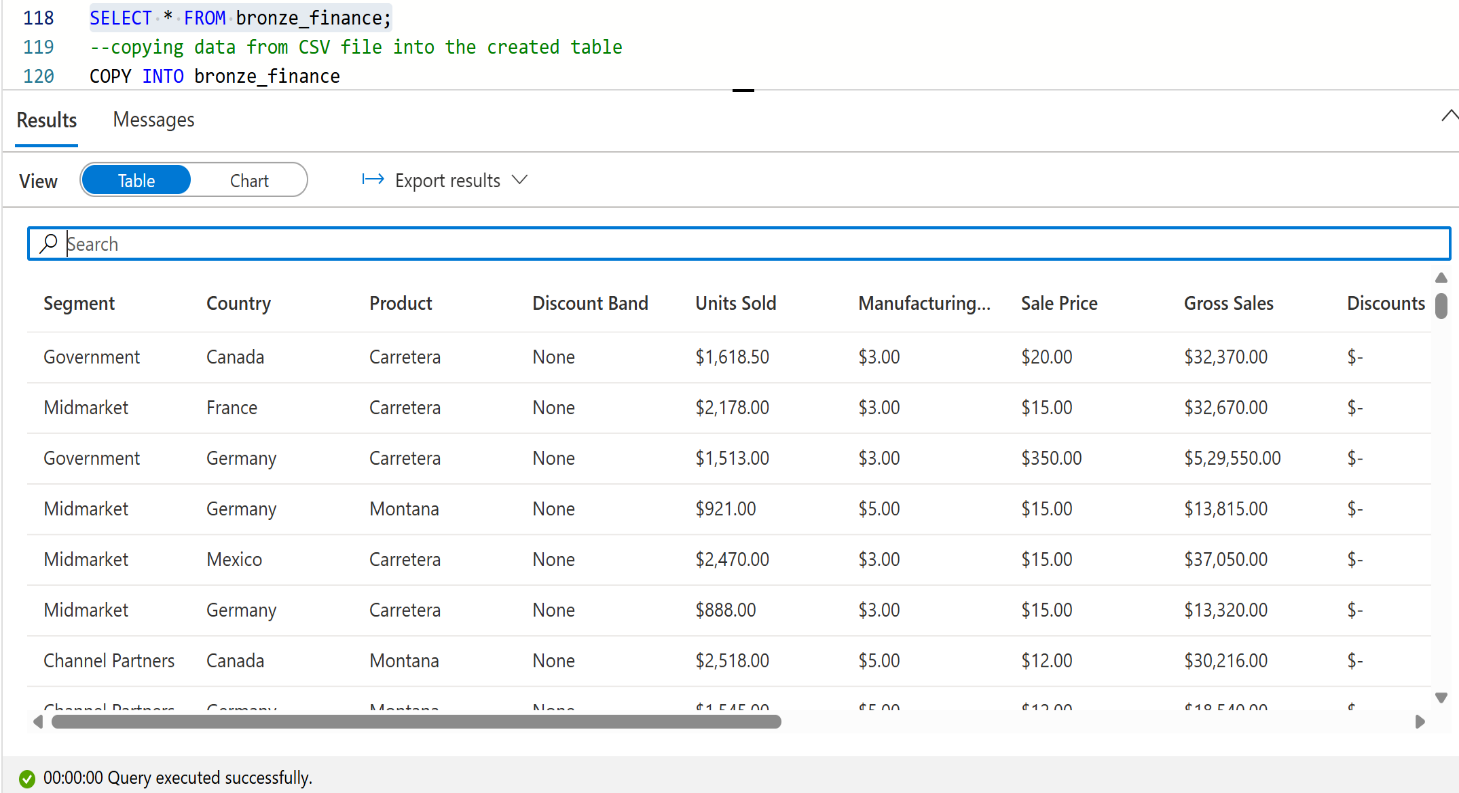
**Created the bronze layer table, initially the table was empty and to get the data in this table used a COPY command in the sql table creation as follows.**

****

****

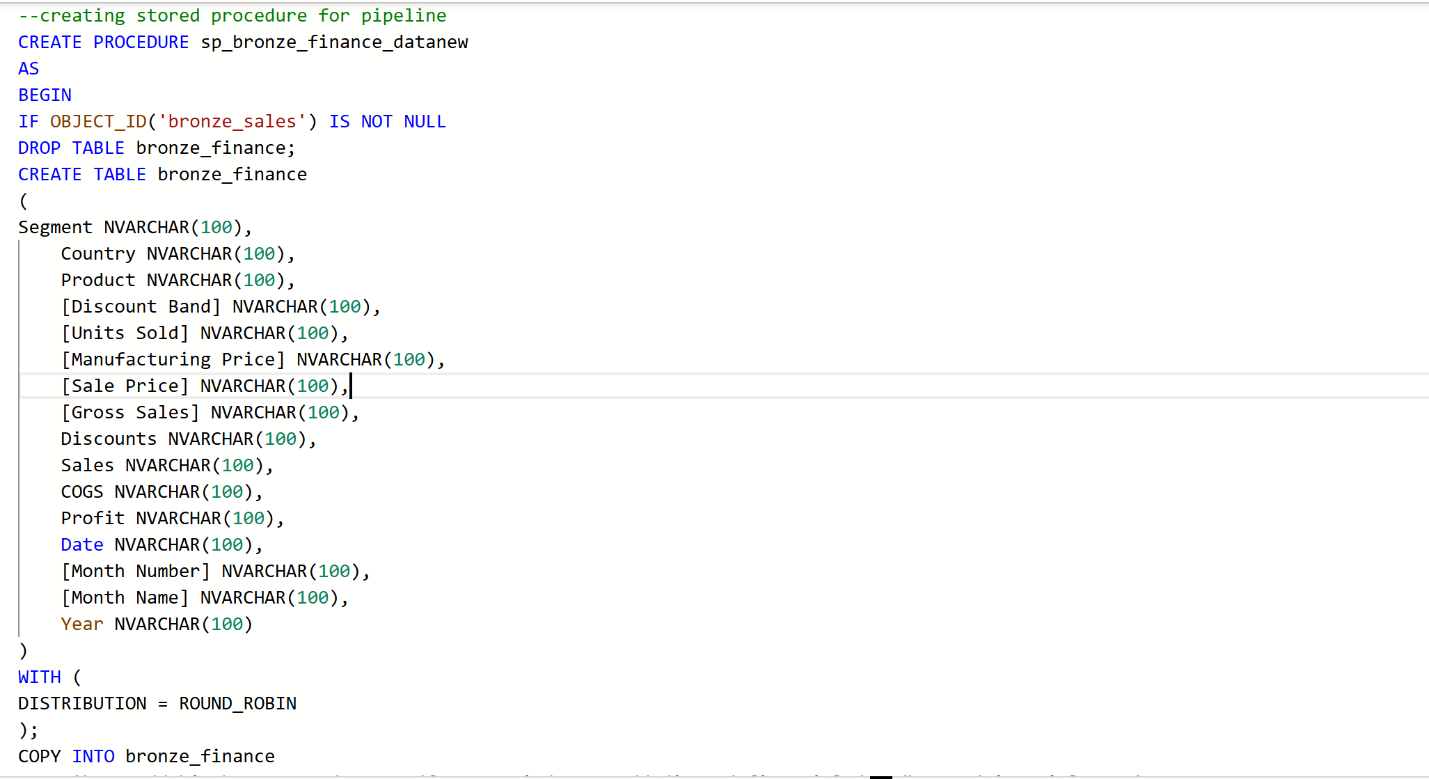
**Again, run the select query then**

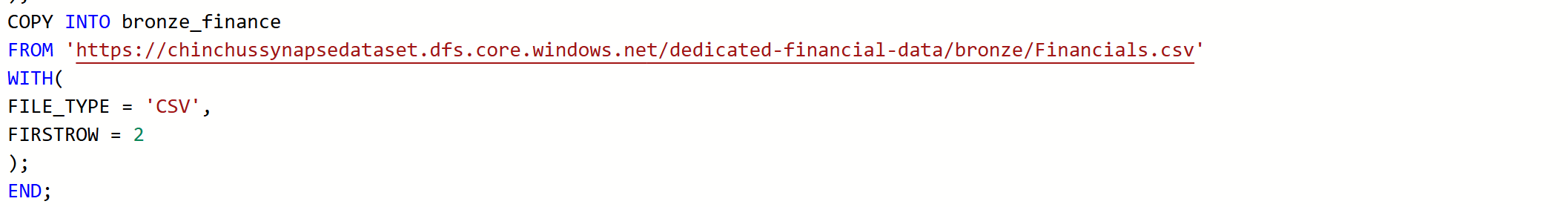
**Result**

****

**Pipeline- Bronze table creation and copy data to it**

**Created a Stored Procedure that includes both tasks - table creation and COPY data**

****

****

**Attach this activity to the current pipeline and debugged it**

A screenshot of a computer

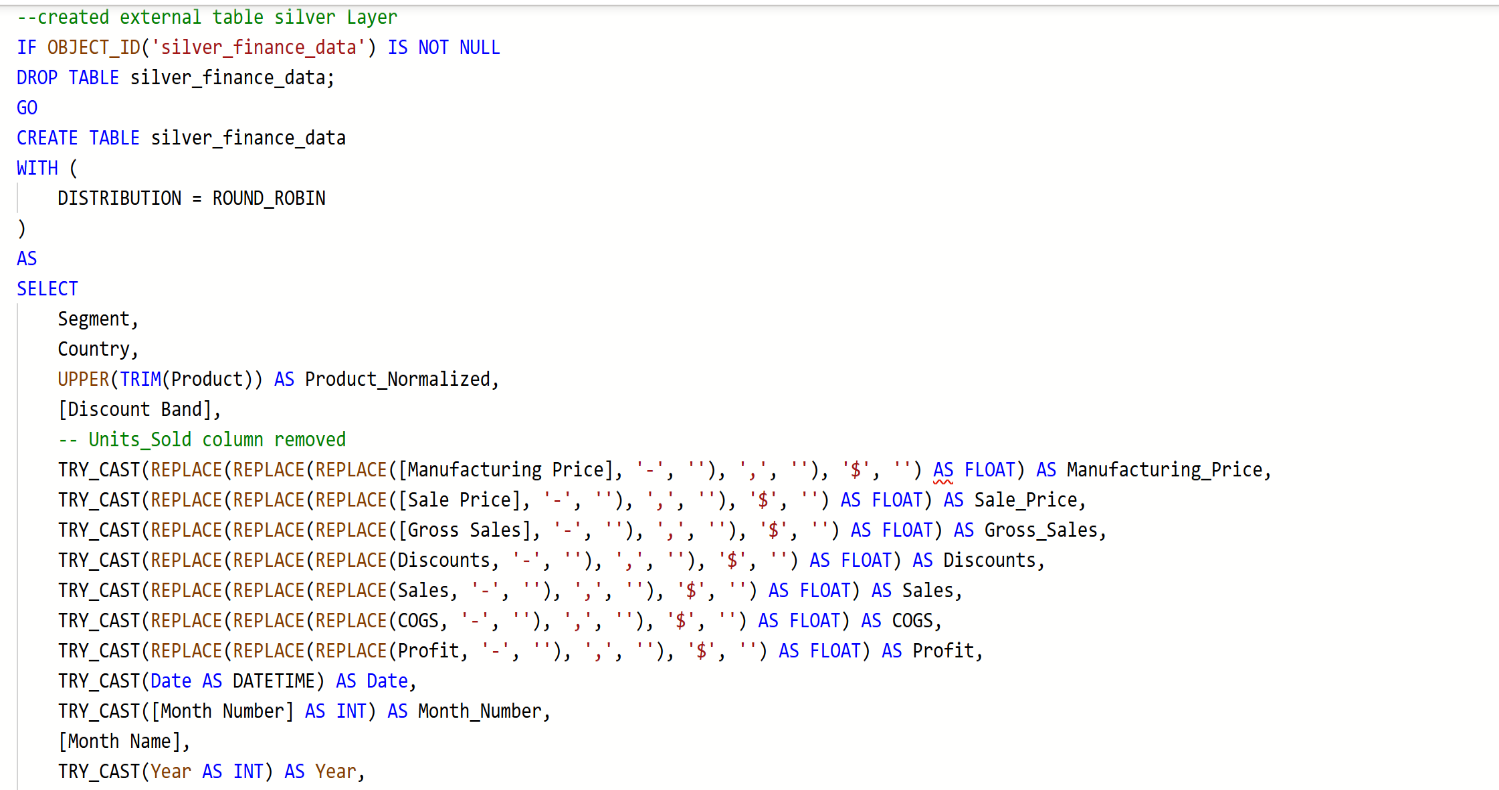
AI-generated content may be incorrect.

**Silver layer creation - Data Cleaning & basic transformations**

From the data viewed in MS Excel I have found that some basic cleaning needed which includes

* Remove '$' sign, “,” (comma) and '-' sign from all columns where they are present
* Change datatype from objects to int after the above two.
* A new column: **Profit Margin = Profit / Sales**
* Normalize product names (e.g., trim and make upper case)
* Remove rows with **Sales = 0** (to avoid divide-by-zero or irrelevant rows)

**Silver table creation**

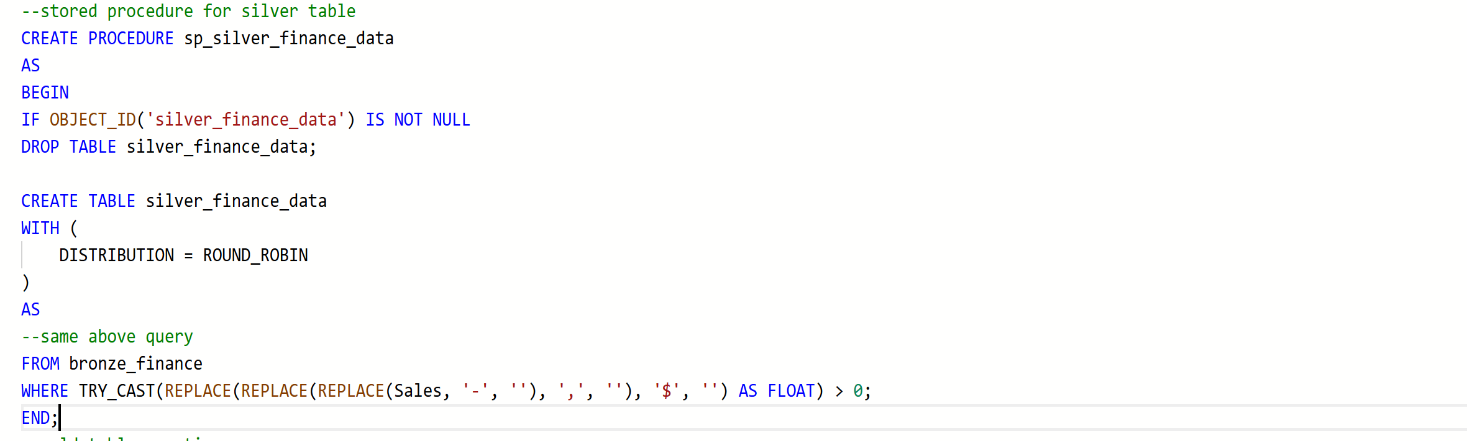
****

**A group of small colorful objects

AI-generated content may be incorrect.**

**Pipeline > Silver table creation**

**Created a Stored Procedure for the above table**

****

**Attached it to the SP activity**

A screenshot of a computer

AI-generated content may be incorrect.

**Debug the pipeline**

A screenshot of a computer

AI-generated content may be incorrect.

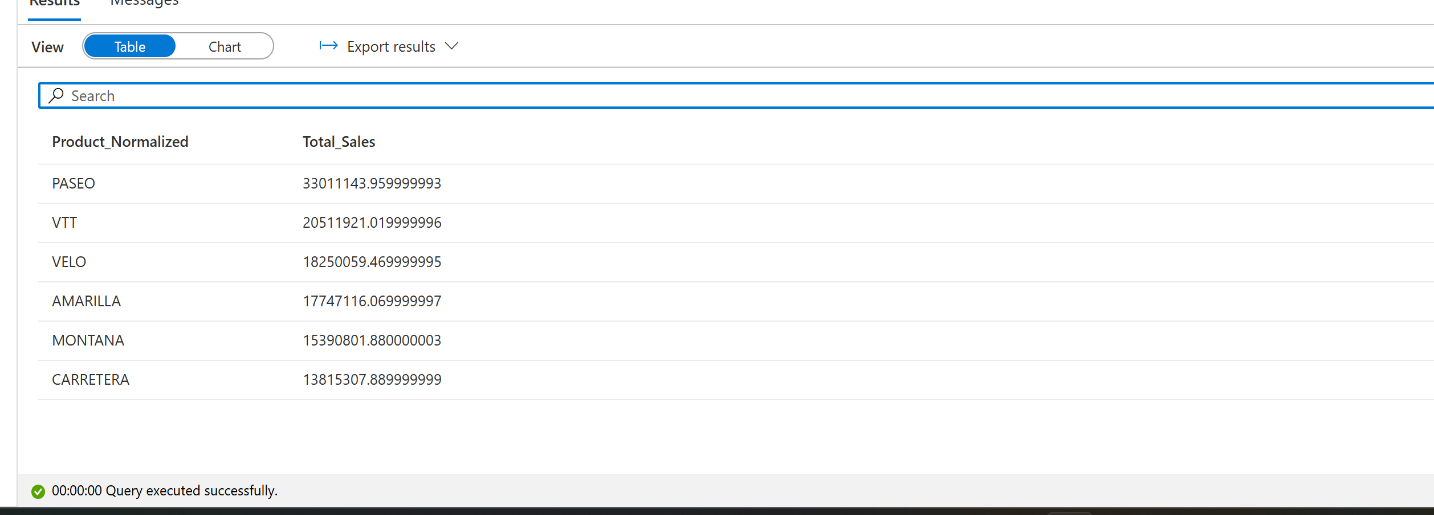
**Creating Gold layer Table**

**This incudes 3 tables**

****

****

**Result:**

****

**A screenshot of a computer

AI-generated content may be incorrect.**

**A screenshot of a computer

AI-generated content may be incorrect.**

**Pipeline Creation**

**Created a single stored procedure for the Gold layer including all the tables.**

**1 Attached a stored procedure activity**

A screenshot of a computer

AI-generated content may be incorrect.

**Debug the pipeline**

A screenshot of a computer

AI-generated content may be incorrect.

**Final pipeline looks like**

**A screenshot of a computer

AI-generated content may be incorrect.**

**Project Summary**

* Data Storage: Raw data was ingested from Azure Data Lake Storage (ADLS) into the Bronze Layer (CSV format).
* Silver Layer: Cleaned and normalized data was stored as Parquet files.
* Gold Layer: Business logic and aggregated views were created using both Dedicated and Serverless SQL Pools.
* Visualization: Power BI was connected to Synapse SQL endpoints for reporting.

**Comparison: Serverless and Dedicated SQL pool**

* **Performance:** Serverless, Optimal for ad-hoc or lightweight queries and Dedicated, High performance for complex joins.
* **Cost**: Pay-per-query (cost-effective) for Serverless and High (provisioned, continuous) for Dedicated.
* **Use Case**: Serverless is great for on-demand, exploratory use where Dedicated is Best for batch loads, heavy BI needs.

**Conclusion**

* This project demonstrated the power of a Medallion architecture in Synapse combined with Power BI visualization to deliver actionable insights.
* Comparing Dedicated vs Serverless SQL Pools revealed trade-offs between performance and cost, suggesting a hybrid model where Serverless is used for exploration and reporting, while Dedicated handles ETL and heavy analytics.