## Retail Sales Data Pipeline Using Azure Medallion Architecture

### **Introduction**

In the modern data-driven world, organizations generate vast amounts of data daily. To efficiently manage, transform, and derive insights from this data, a robust data pipeline is essential. This project focuses on building an automated Retail Sales Data Pipeline using Azure Data Factory (ADF), Azure Databricks, and Azure Blob Storage, following the Medallion Architecture principles.

The project aimed to:

* Ingest raw sales data and store it in a centralized Bronze Layer.
* Transform and clean the data into a structured format in the Silver Layer.
* Aggregate and summarize the data into Gold Layer for analytics and insights.
* Apply machine learning (Linear Regression) for sales forecasting.
* Automate the entire process to ensure seamless workflows with minimal manual intervention.

### **Objectives**

The primary objectives of the project were:

1. Ingest Data: Load raw sales data (CSV) into the Bronze Layer.
2. Clean and Transform Data: Process raw data to remove inconsistencies and store it as Parquet files in the Silver Layer.
3. Aggregate Data: Generate region-wise, product-wise, price-wise and date-wise summaries in the Gold Layer.
4. Forecast Sales: Use Linear Regression to predict sales trends for better decision-making.
5. Automation: Design ADF pipelines to automate the data ingestion, transformation, and aggregation workflows.
6. CI/CD Integration: Use Azure DevOps for seamless deployment and version control.

### **System Design**

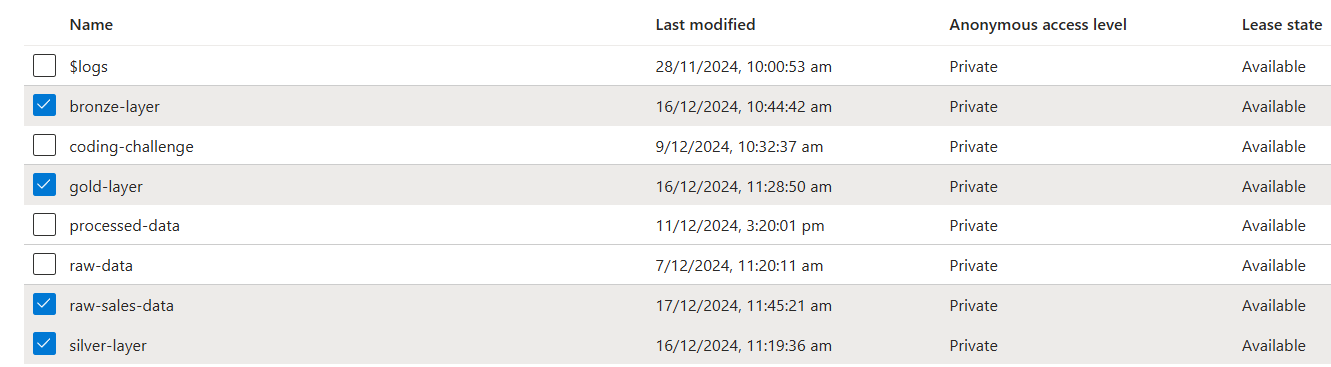
The solution consisted of:

* Data Ingestion: Copy raw data into the Bronze Layer using Azure Data Factory.
* Data Transformation: Clean and transform the data in Azure Databricks.
* Data Aggregation: Summarize and analyze the data in Databricks Notebooks.
* CI/CD Integration: Use Azure DevOps Pipelines for deployment and management.

### Folder Structure

The following folder structure was used across Azure Blob Storage:

* raw-sales-data (Hot Storage): Contains unprocessed raw data in CSV format.
* bronze-layer: Stores raw data in Parquet format for staging.
* silver-layer: Contains cleaned and transformed data.
* gold-layer: Stores aggregated summaries.

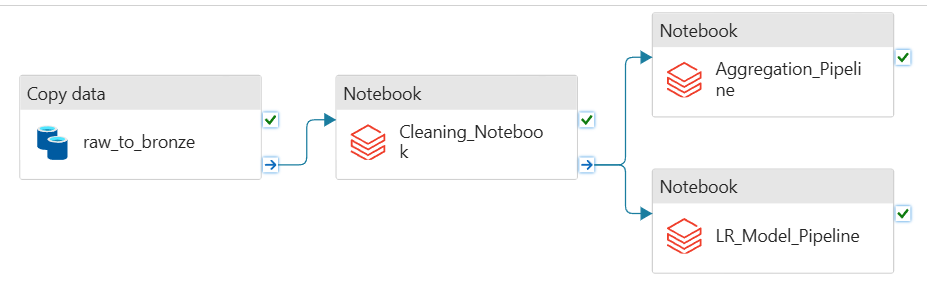


This segregation follows the Medallion Architecture and ensures optimized performance and cost efficiency.

### **Implementation**

The solution was implemented using a pipeline in Azure Data Factory that orchestrated the execution of three Databricks notebooks for data transformation and aggregation. The pipeline also included data ingestion activities.

#### **Pipeline Overview**



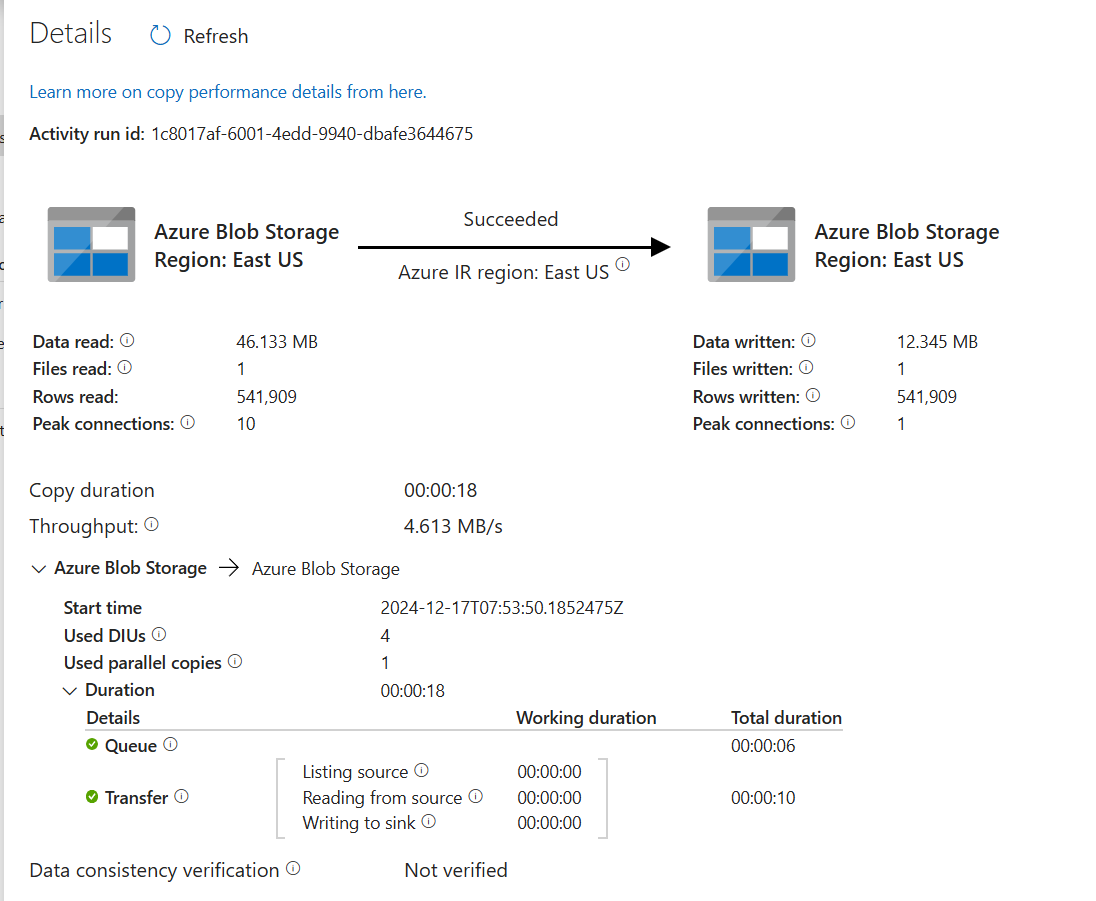
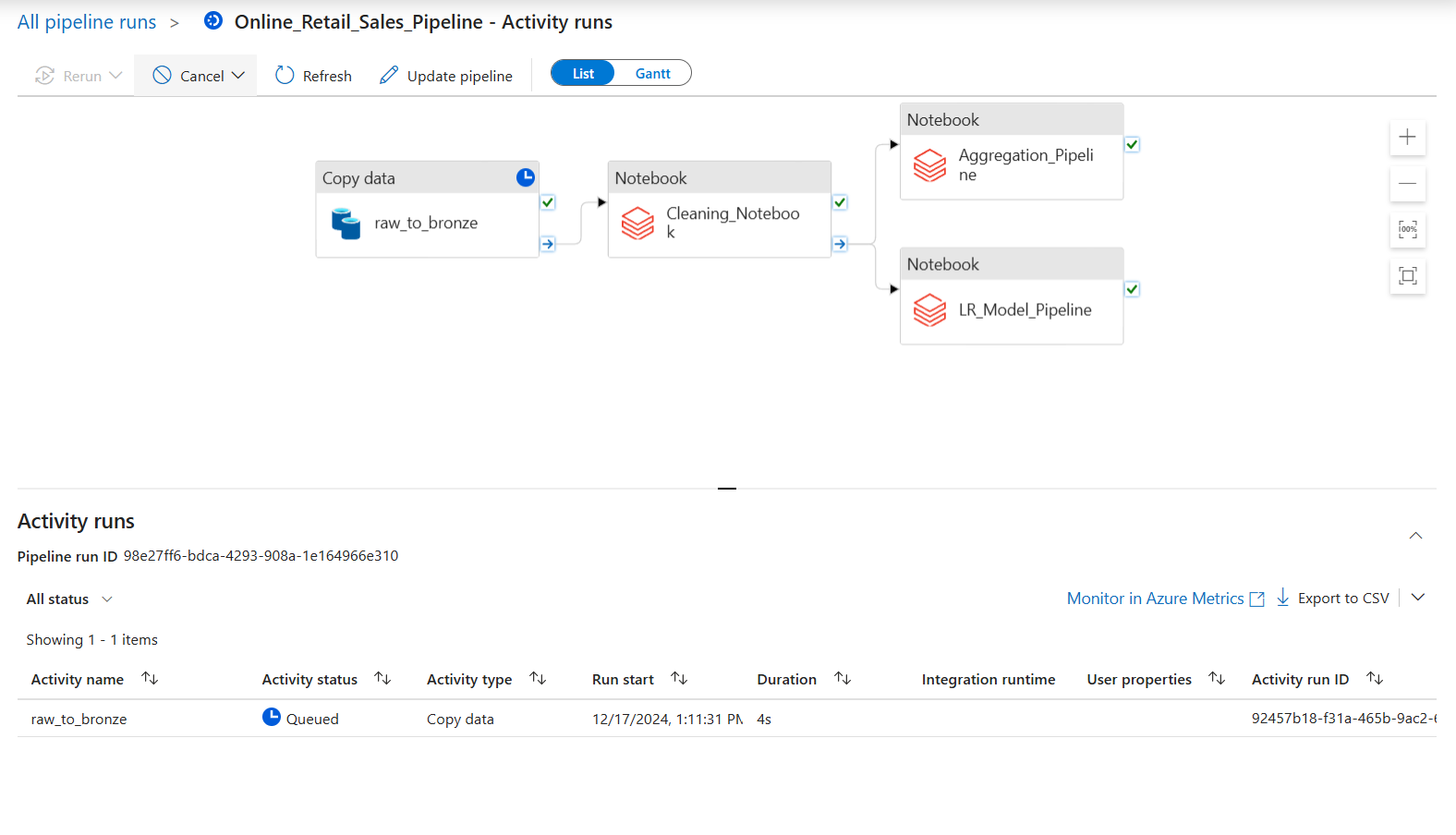
The pipeline was designed to execute the following steps sequentially:

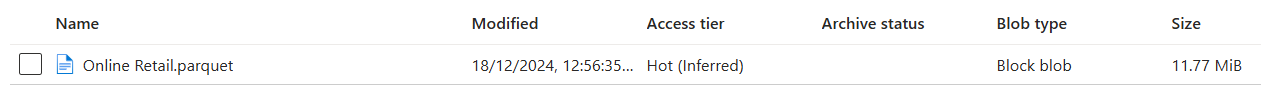
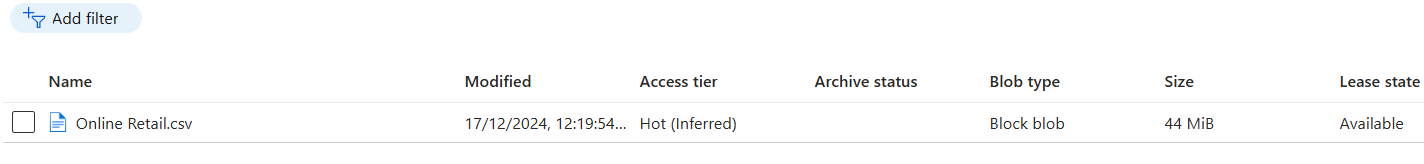
1. Ingest Raw Data (ADF Copy Activity).
2. Transform and Clean Data (Databricks Notebook 1).
3. Generate Aggregated Summaries (Databricks Notebook 2).
4. Apply Machine Learning for Sales Forecasting (Databricks Notebook 3).

### **Key Components**

#### **1. Data Ingestion**

* Tool: Azure Data Factory (ADF)
* Process:
  + Raw sales data (CSV files) from the raw-sales-data container was ingested into the Bronze Layer.
  + Data was stored as Parquet files for optimized performance.

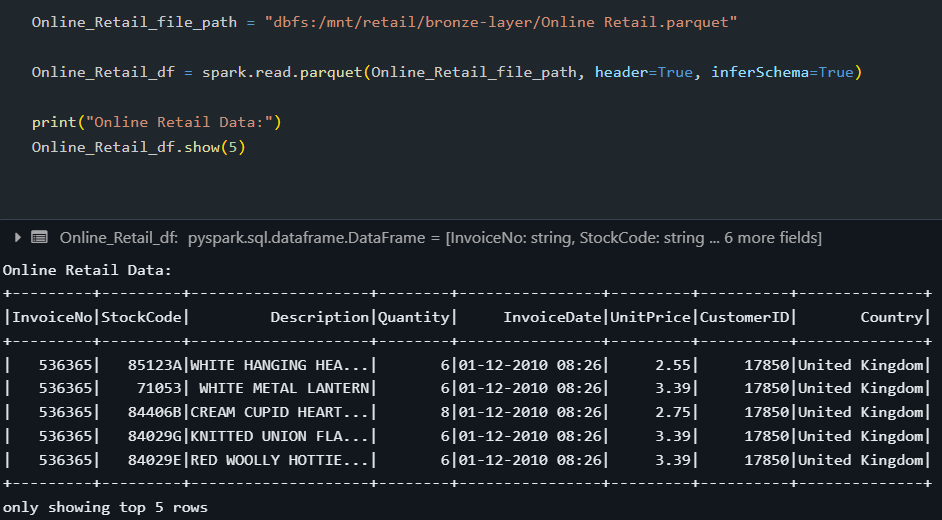




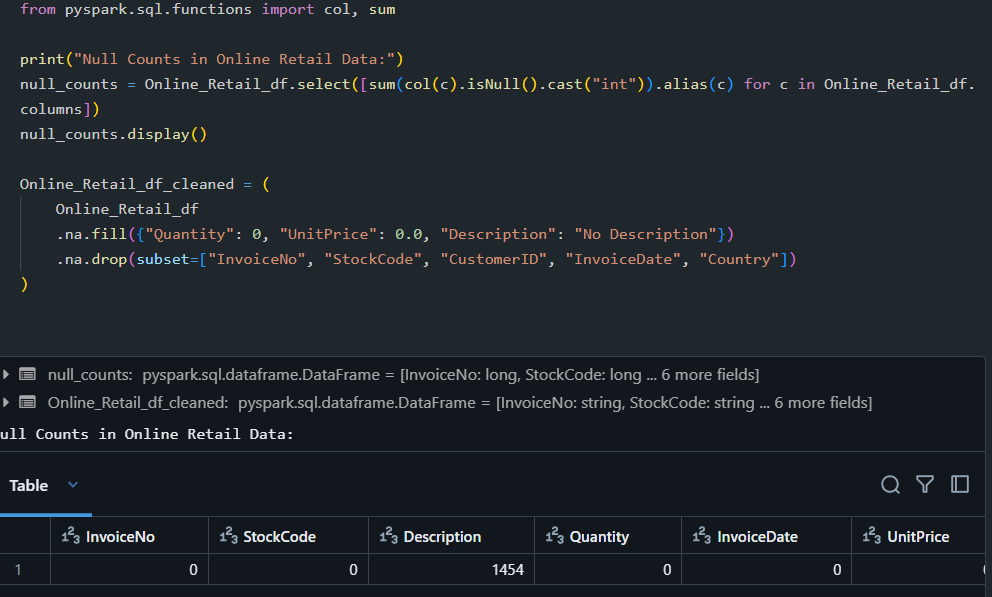
**2. Data Transformation**

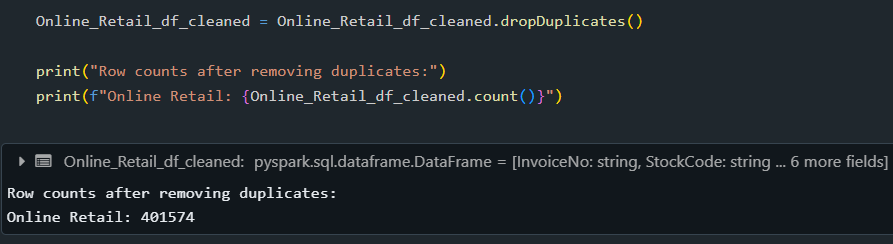
The transformation process cleaned and processed the raw data to ensure quality and consistency.

* Notebook 1: Retail Sales Data Cleaning
  + Tool: Azure Databricks (PySpark)
  + Key Steps:
    - Read raw Parquet data from the Bronze Layer.

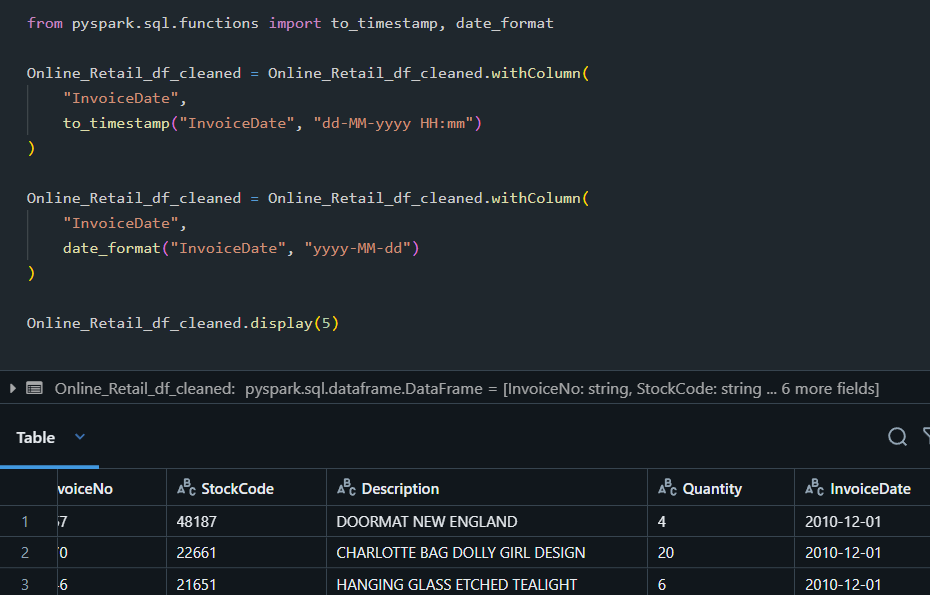


* + - Removed null values and duplicate records.

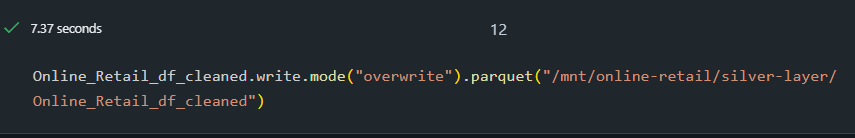




* + - Standardized date formats.



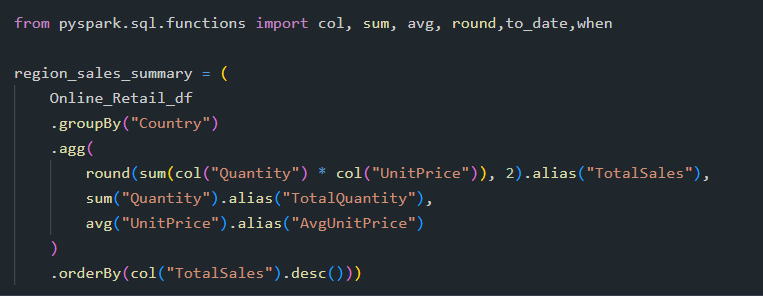
* + - Stored the cleaned data in the Silver Layer as Parquet files.

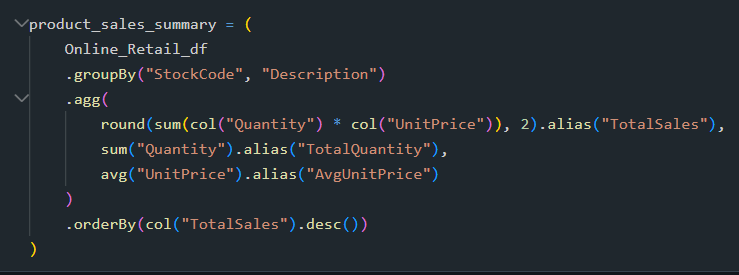


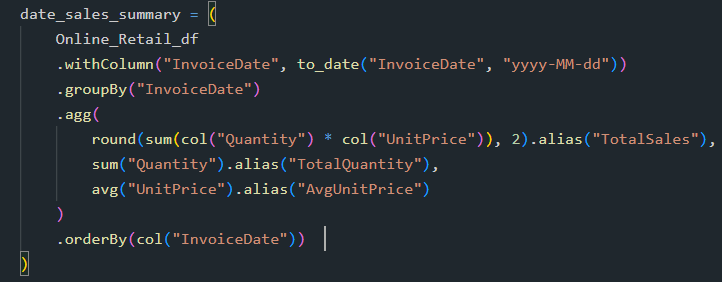
#### **3. Data Aggregation**

Aggregated summaries were generated to derive insights into retail sales trends.

* Notebook 2: Retail Sales Data Aggregation
  + Tool: Azure Databricks (PySpark)
  + Key Steps:
    - Created region-wise, product-wise, date-wise, and price summaries.

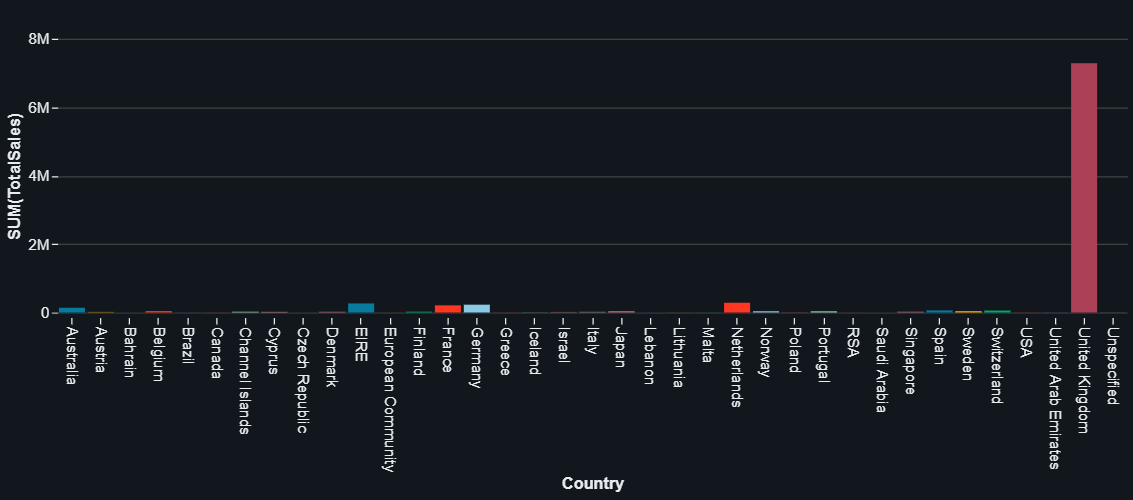


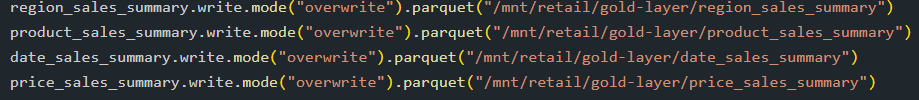




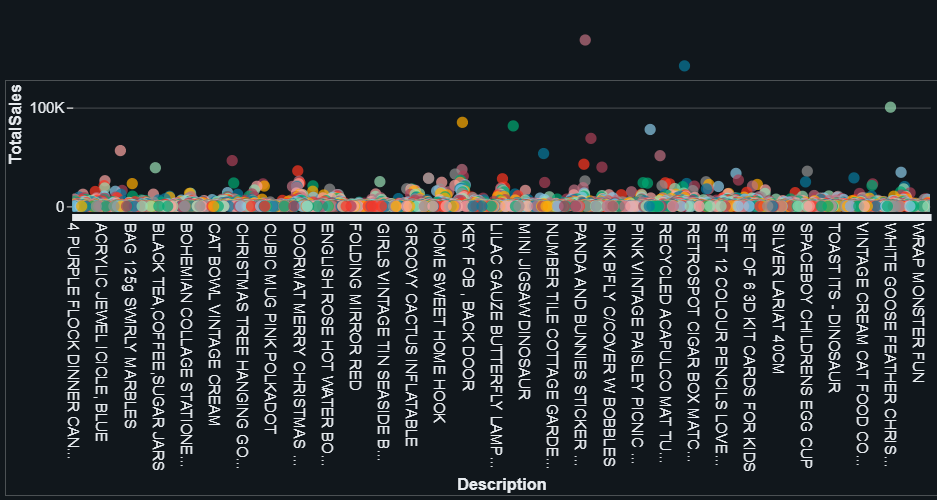


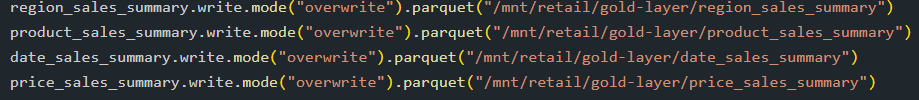
* + - Output data was saved in the Gold Layer:
      * Region Summary: /mnt/retail/gold-layer/region\_sales\_summary



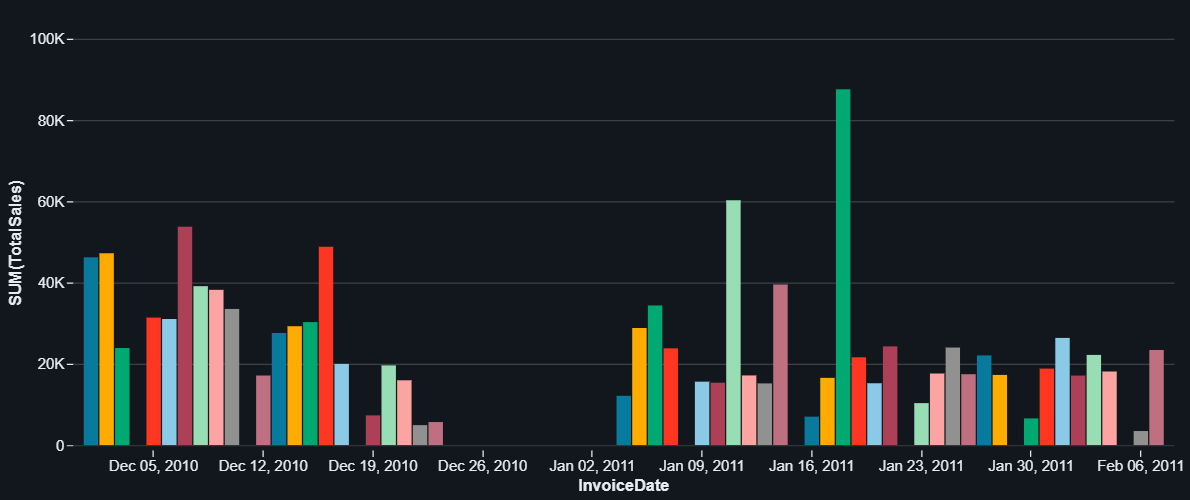


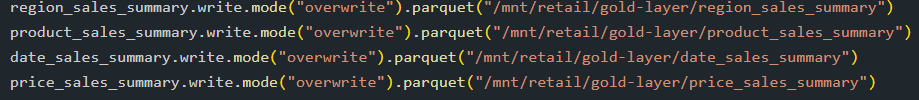
* + - * Product Summary: /mnt/retail/gold-layer/product\_sales\_summary



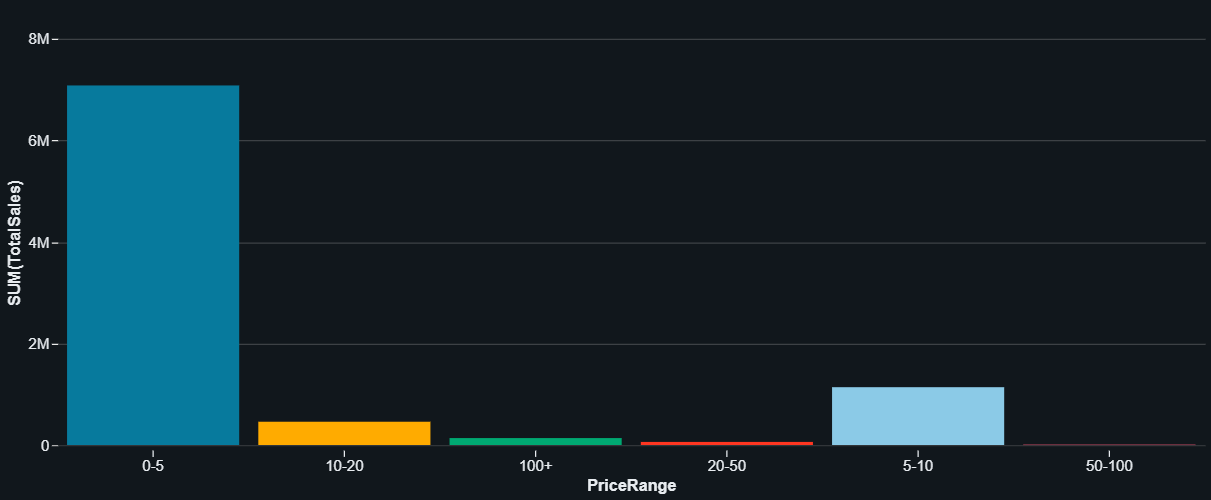


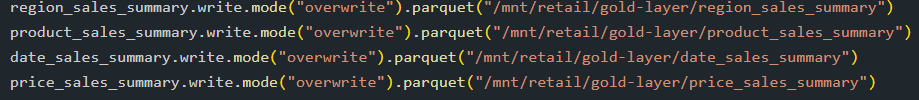
* + - * Date Summary: /mnt/retail/gold-layer/date\_sales\_summary





* + - * Price Summary: /mnt/retail/gold-layer/price\_sales\_summary



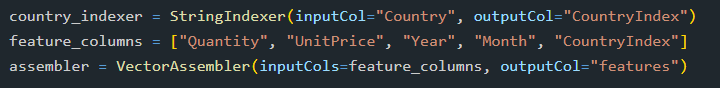


#### **4. Sales Forecasting**

To predict future sales trends, a Linear Regression Model was applied:

* Notebook 3: Retail Sales Data Model LR
  + Tool: Azure Databricks
  + Trained the Linear Regression model using historical sales data.

Feature Engineering



Splitting Data



Linear Regression Model



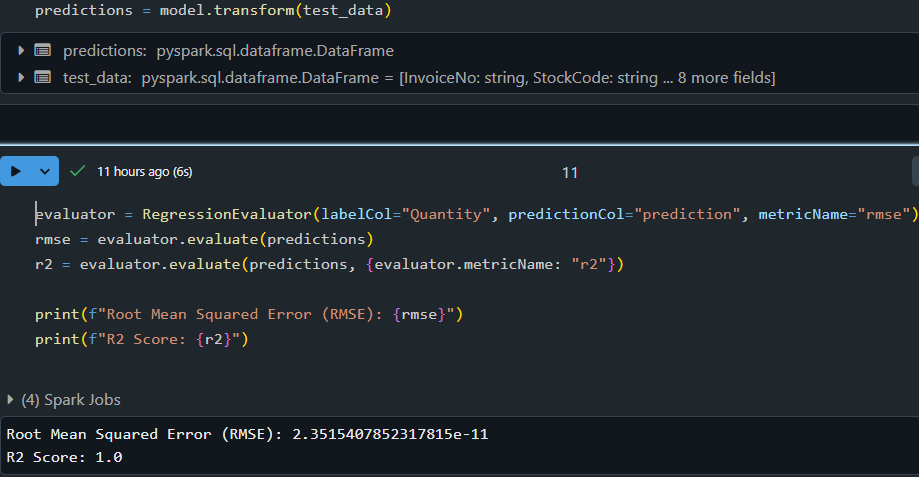
Pipeline Construction



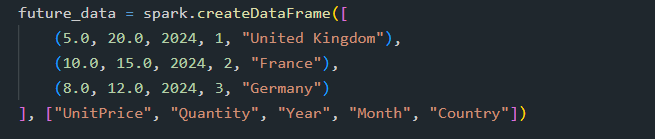
Training Model



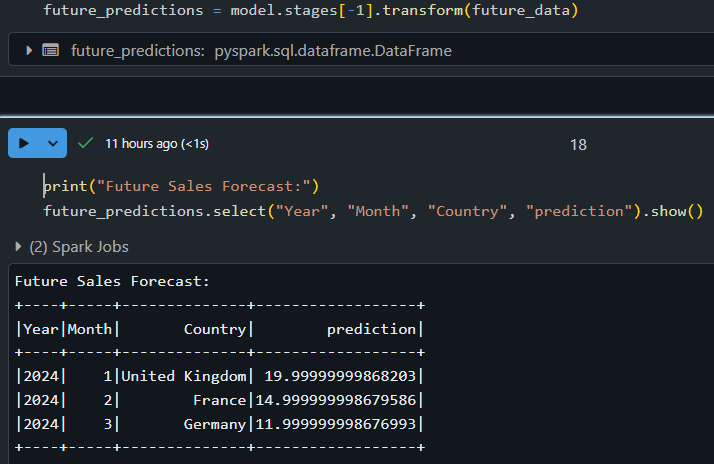
Transforming Test Data and Evaluating RMSE and Rsquare



Simulating Data for Future Dates



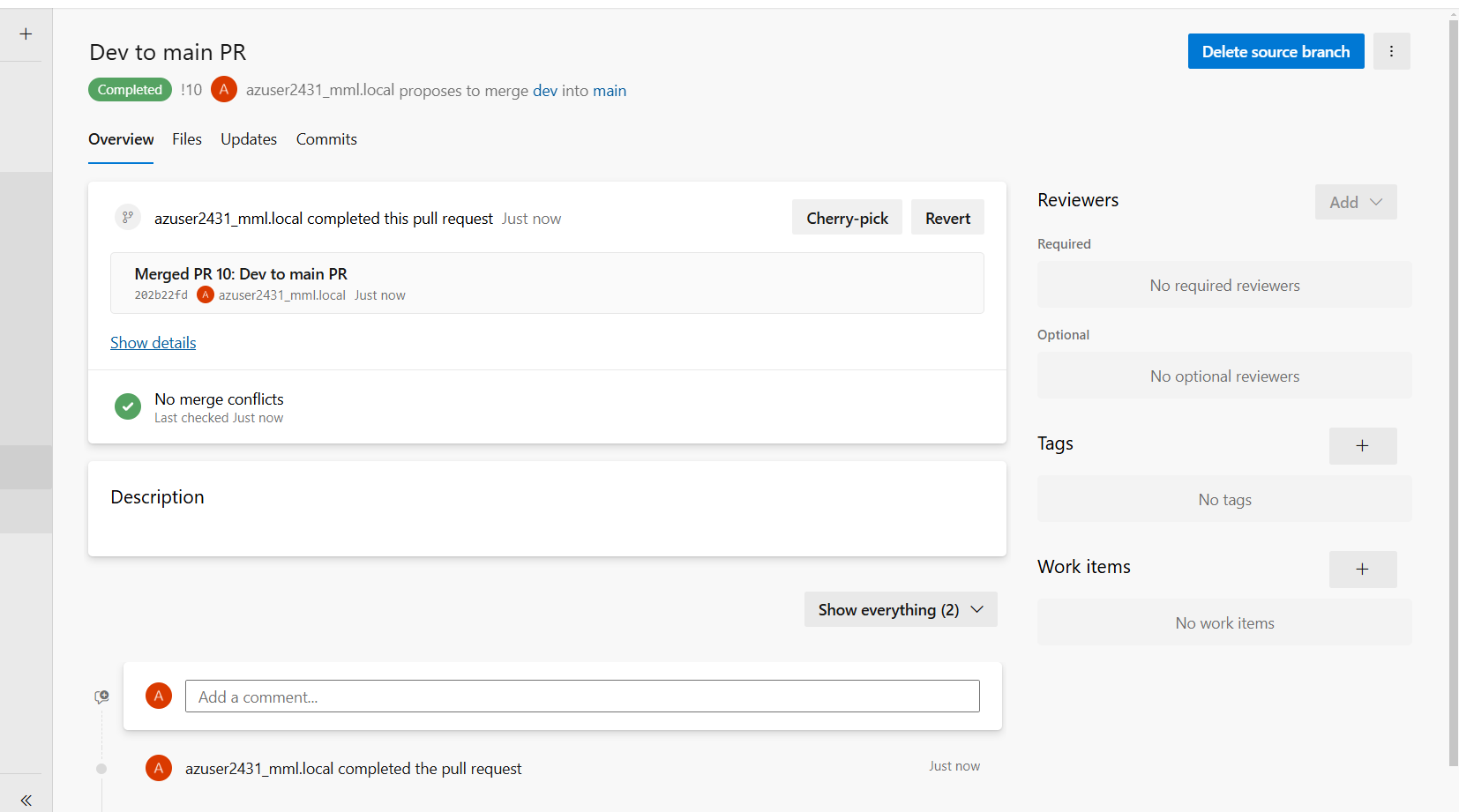
After all the transformation from trained pipeline to future data making prediction

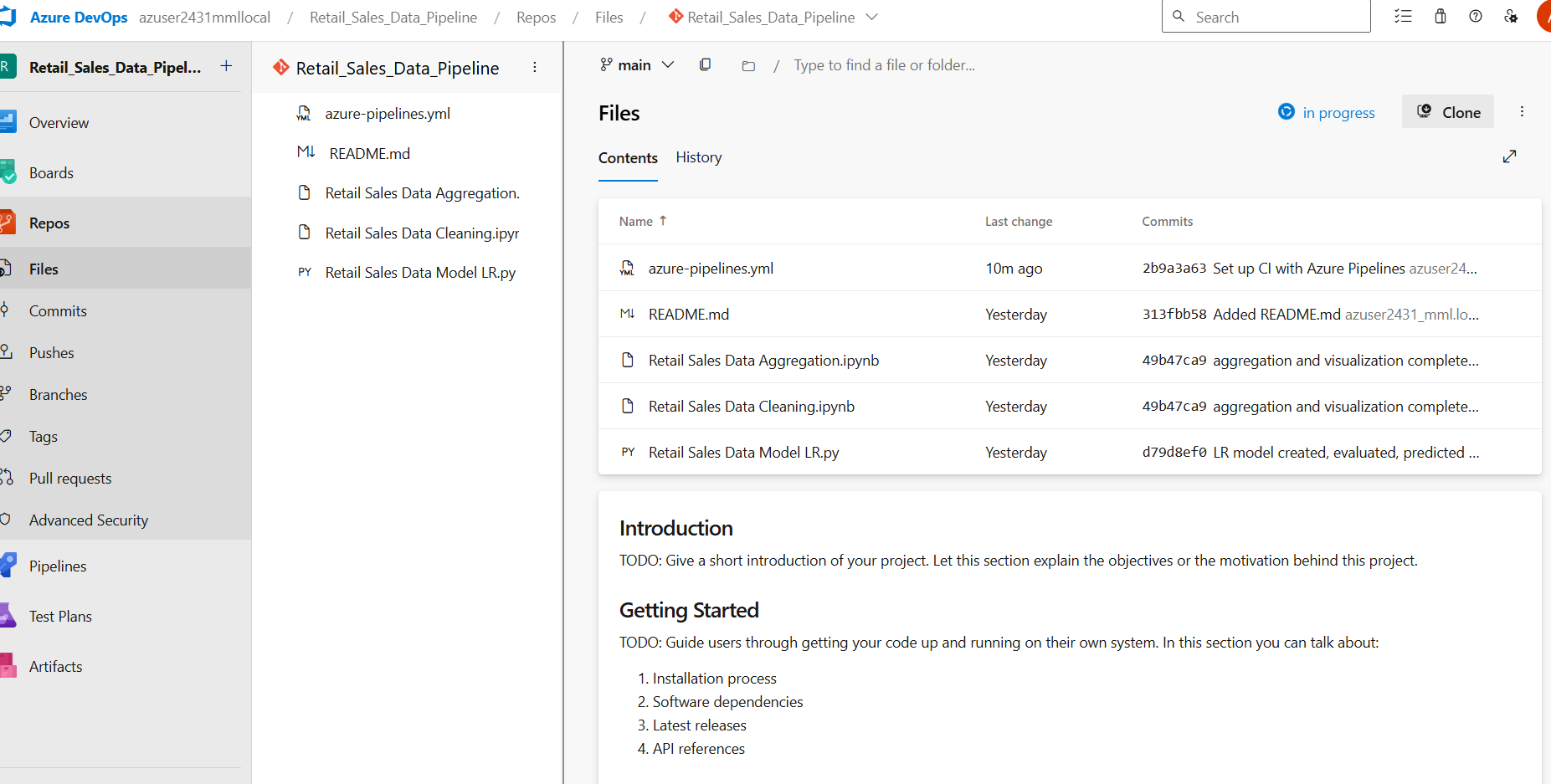
We can see **Future Sales Forecast using Linear Regression Model**

### **CI/CD Pipeline**

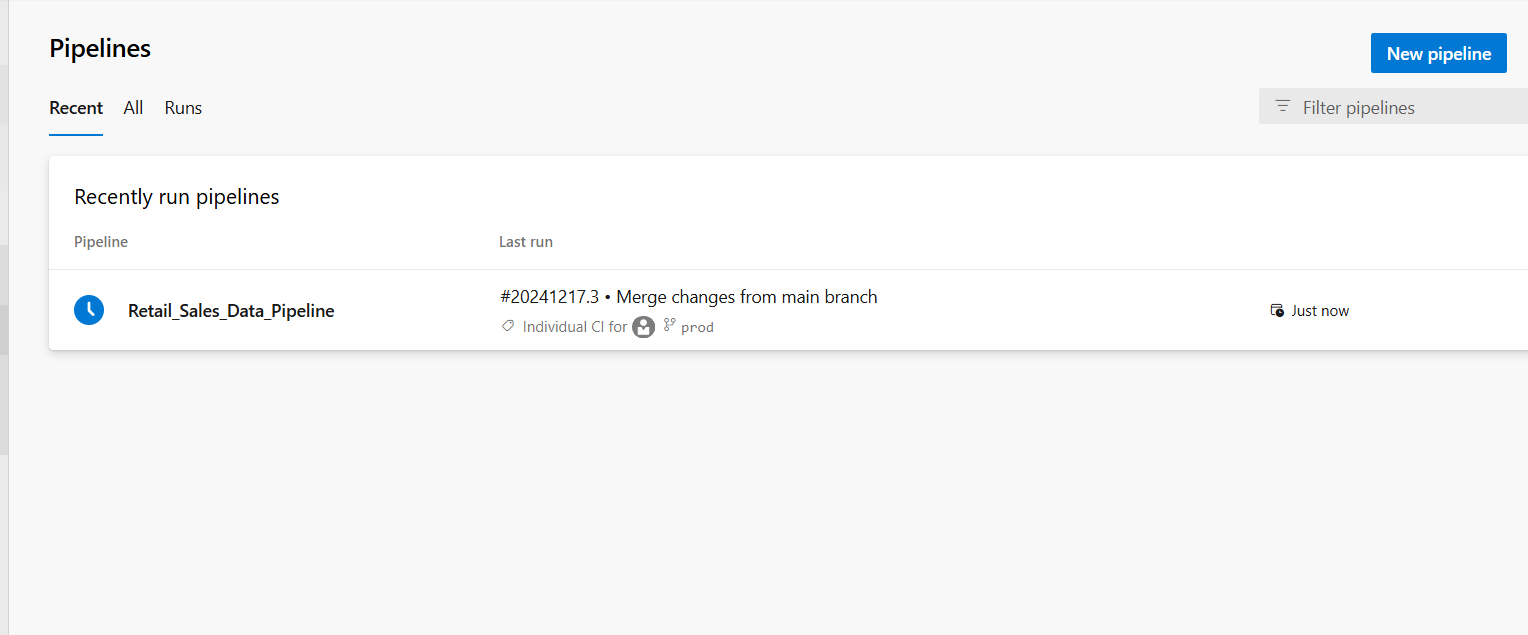
To ensure smooth development and deployment, Azure DevOps was used for CI/CD.

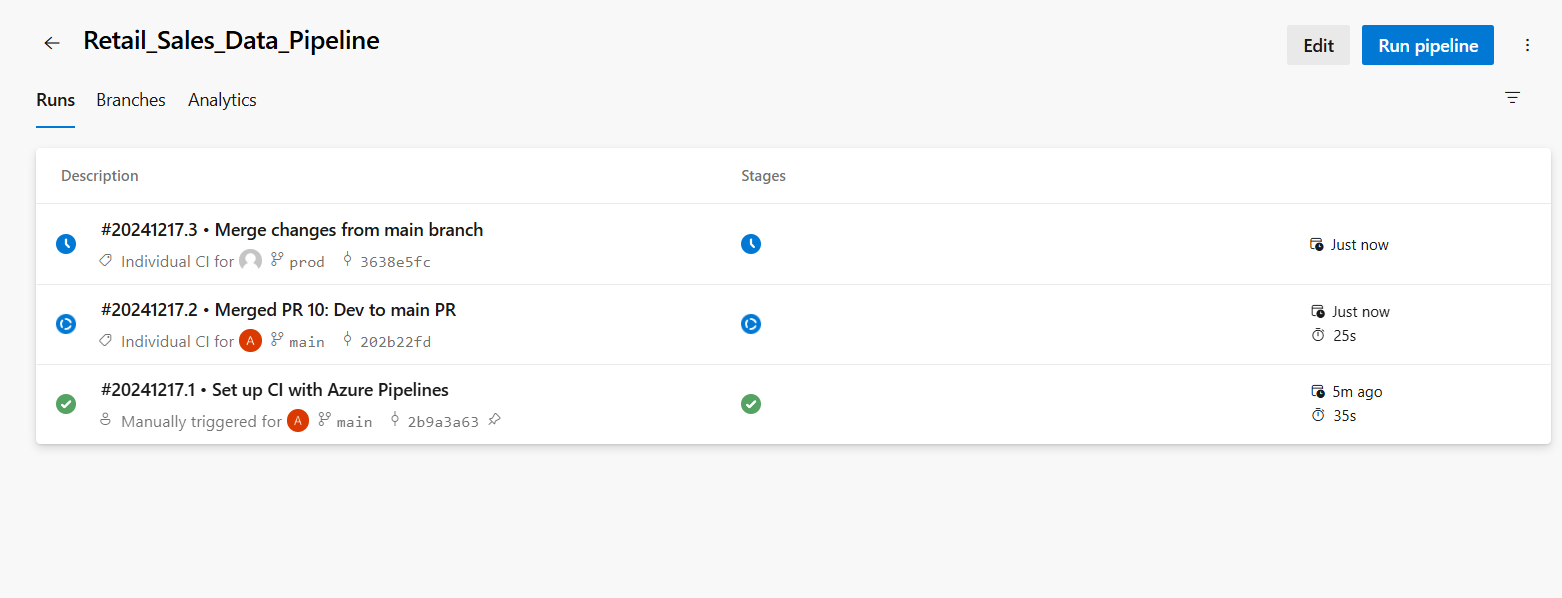
1. Branches:
   * dev: For development and testing.
   * prod: For production-ready changes.
   * Merged into main for release.

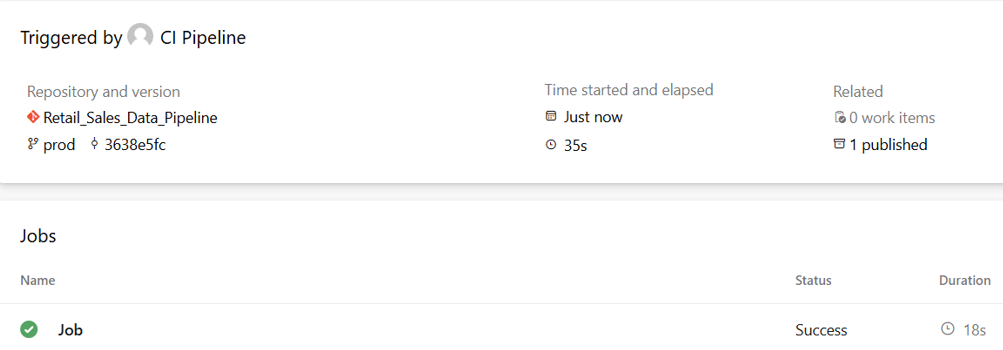


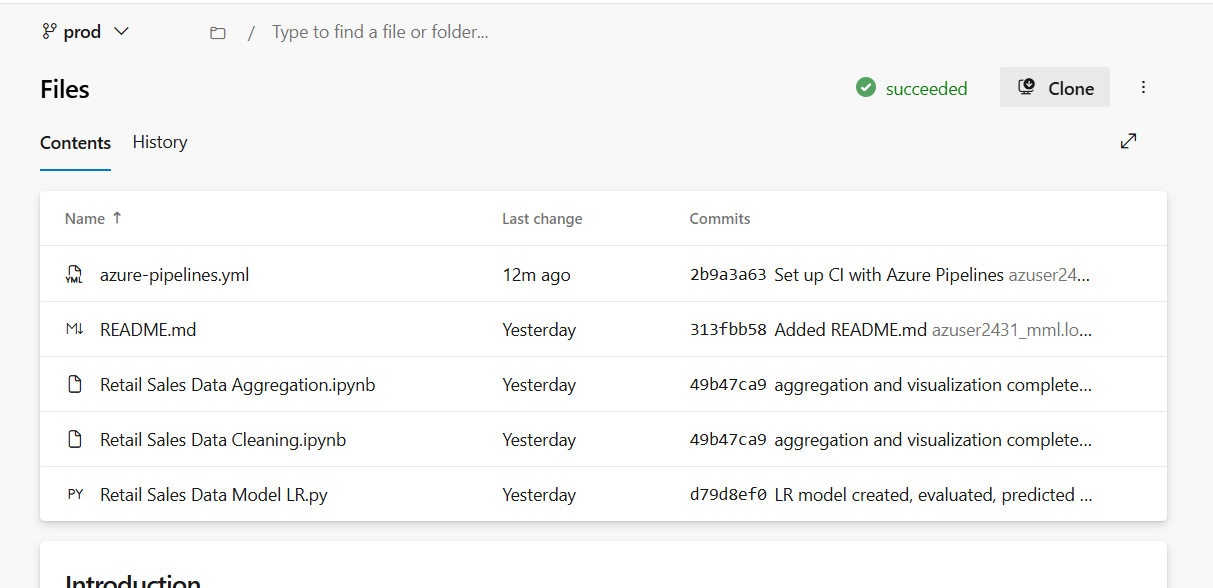


1. Pipeline:
   * Automated deployment of Databricks Notebooks.

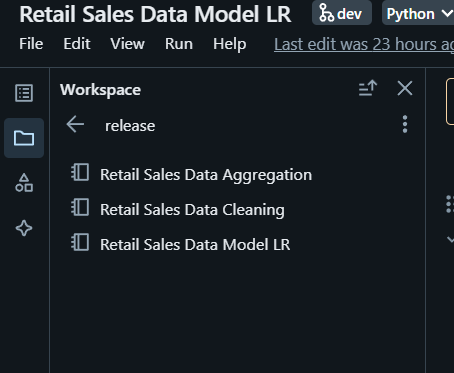




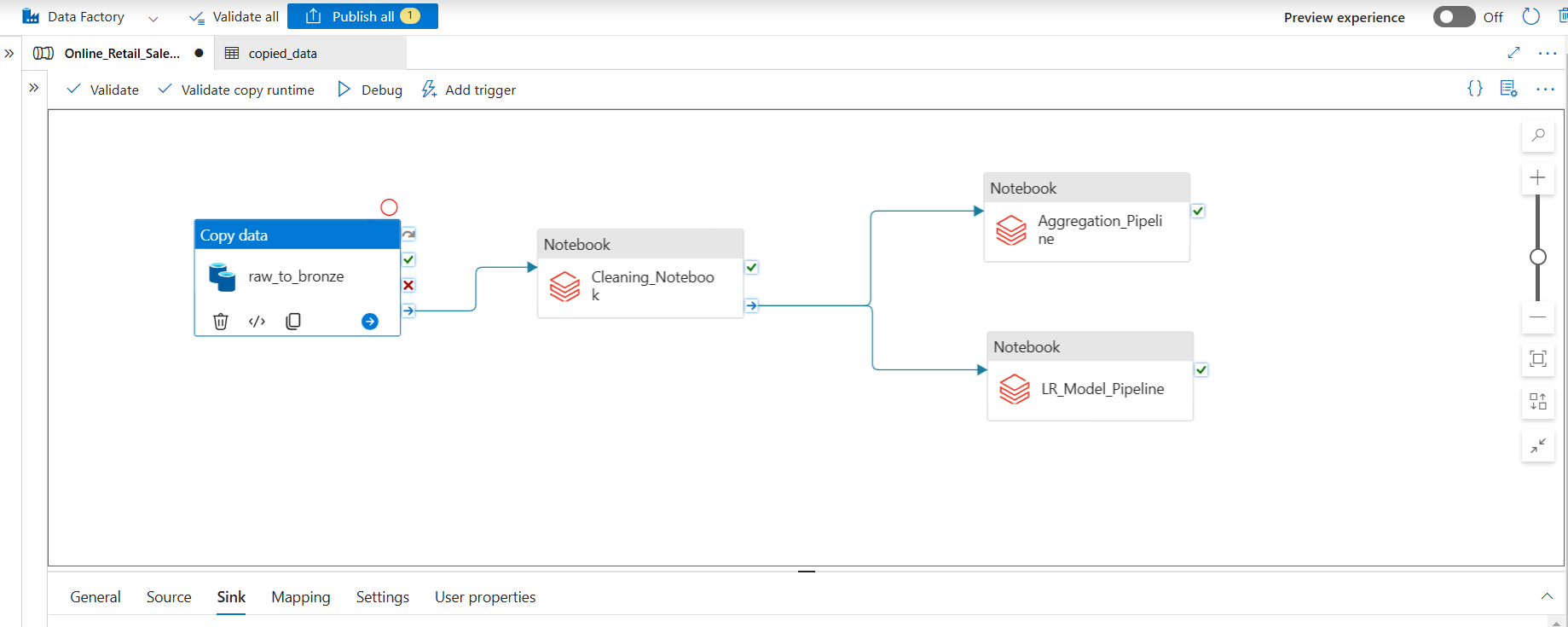




* + Release folders created in the Azure Workspace for final outputs.



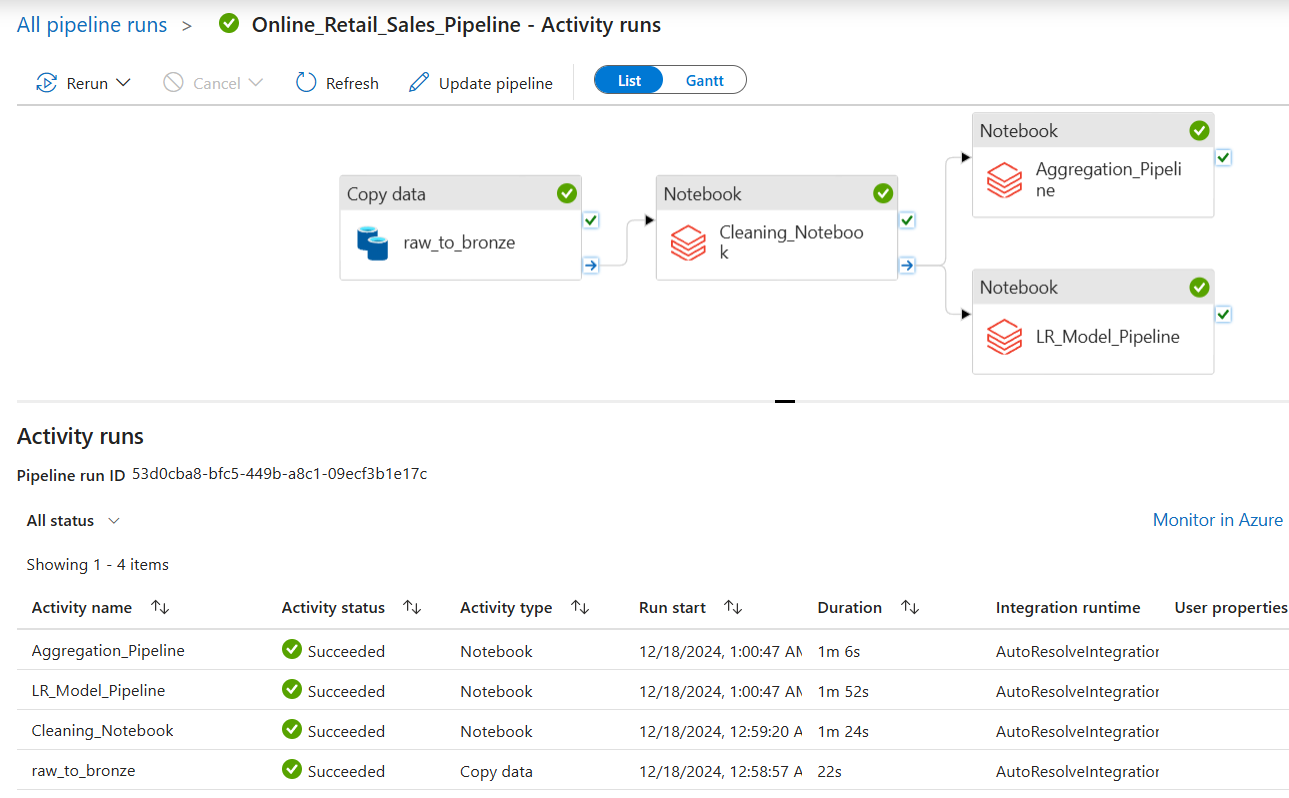
### **Workflow Execution**



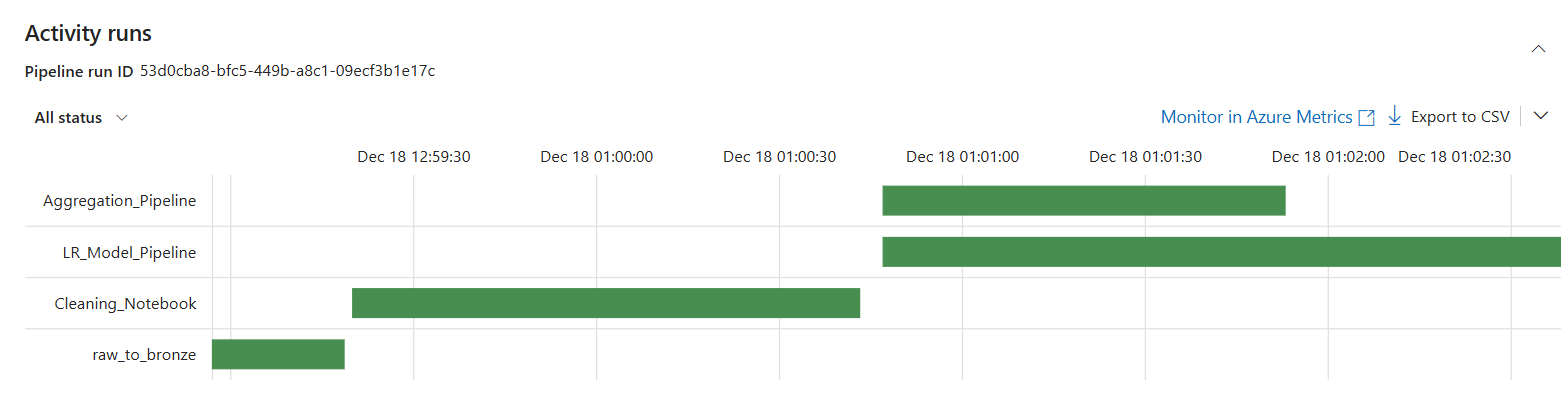
1. Data Ingestion:
   * Raw sales data loaded into the Bronze Layer.
2. Data Cleaning:
   * Data cleaned and standardized in the Silver Layer.
3. Data Aggregation:
   * Summarized data stored in the Gold Layer for analytics.
4. Forecasting:
   * Sales trends forecasted using a Linear Regression model.

### **Validation and Monitoring**

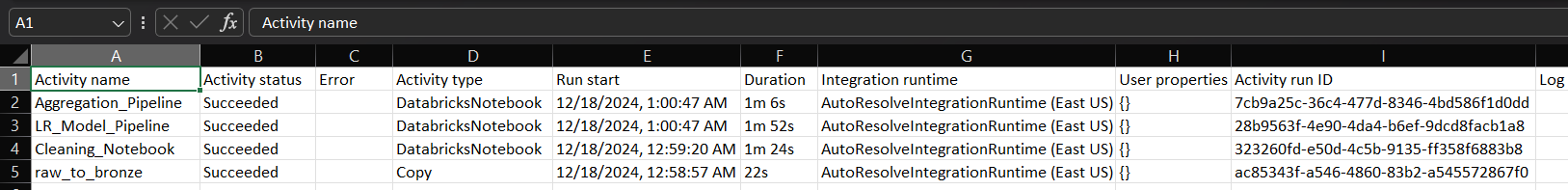
1. Pipeline Monitoring:
   * Monitored pipeline execution in ADF's Monitor section.



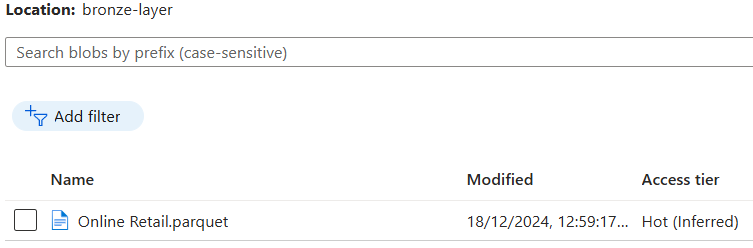
* + Activity runs and statuses were reviewed in the Gantt Chart.

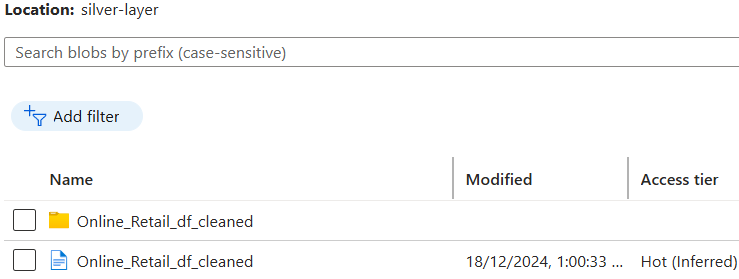


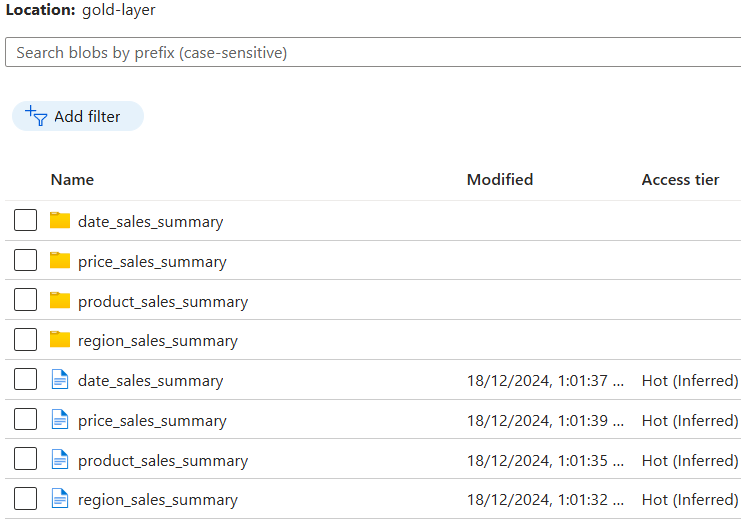
1. Logs:
   * Downloaded activity execution logs as CSV files for audit purposes.



1. Output Verification:
   * Data was verified in the respective storage layers (Bronze, Silver, Gold).







### **Conclusion**

The Retail Sales Data Pipeline was successfully implemented to:

* Automate data ingestion, transformation, and aggregation.
* Generate actionable insights from region-wise and product-wise sales data.
* Forecast sales trends for improved business decisions.
* Implement CI/CD using Azure DevOps for seamless deployment.

This solution demonstrates the scalability and efficiency of Azure services in managing large-scale data workflows following the Medallion Architecture.