End-to-End Data Ingestion and Transformation through Azure Data Engineering

A Real-World Data Pipeline Project

Presented by:
Arpan Ray, Mainak Mukherjee, Trishita Mukherjee

TABLE OF CONTENT

INTRODUCTION

This project showcases an end-to-end data engineering pipeline using Microsoft Azure, Databricks, and PySpark. Data from multiple sources is ingested through Azure Data Factory, then cleaned, transformed, and processed in Azure Databricks using PySpark. The processed data is stored in Delta Lake, ensuring optimized performance and reliability. By leveraging the orchestration and automation features of ADF, the pipeline achieves scalability, security, and efficiency, making it a robust solution for modern data analytics and business intelligence needs.

PROJECT OVERVIEW

- Build an automated data pipeline using Azure Data Factory and Databricks
- Transform data across Bronze, Silver, and Gold layers

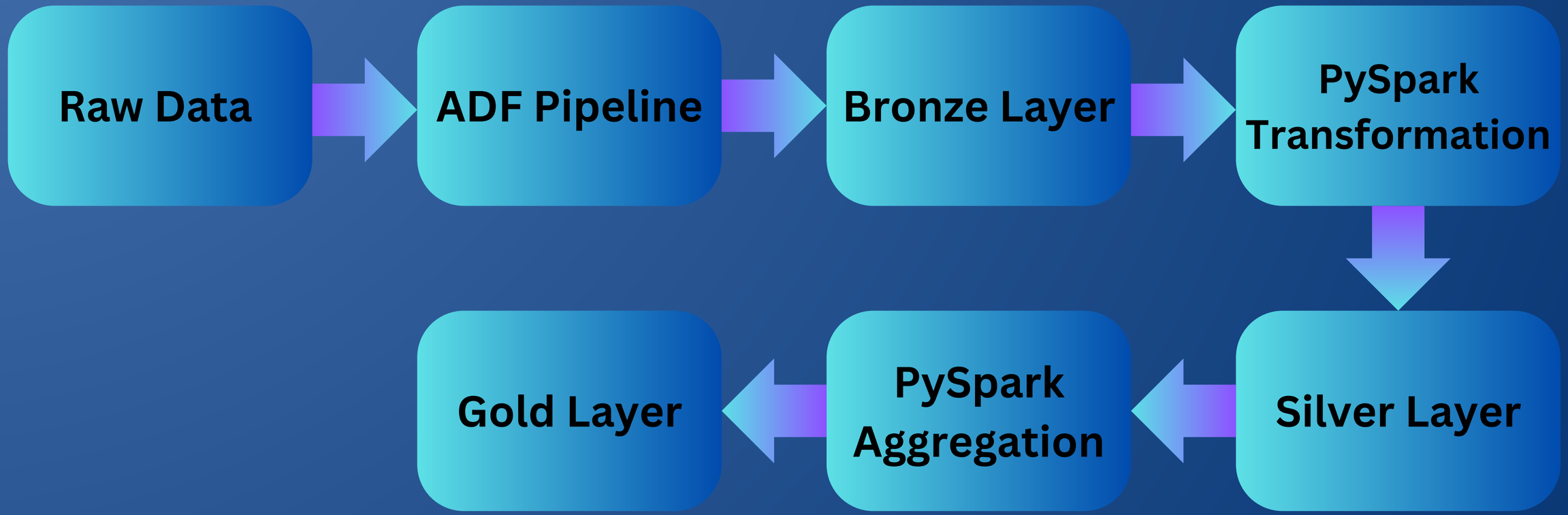
OBJECTIVES

- Build a scalable data pipeline to ingest raw data from diverse sources using Azure Data Factory.
- Transform and clean data efficiently with PySpark in Azure Databricks.
- Optimize data storage and retrieval by leveraging Delta Lake on Azure.
- Automate and orchestrate workflows with ADF for seamless data movement and processing.
- Ensure security, scalability, and reliability across the entire data engineering lifecycle.
- Enable advanced analytics and reporting through well-structured, high-quality data.

TECHNOLOGIES & TOOLS USED

- Azure Data Factory (ADF) - Data ingestion & orchestration
- Azure Databricks - Transformation with PySpark
- Delta Lake - Optimized storage
- Azure Storage - Raw, Bronze, Silver, Gold containers
- Azure AD - Security and Role Assignment

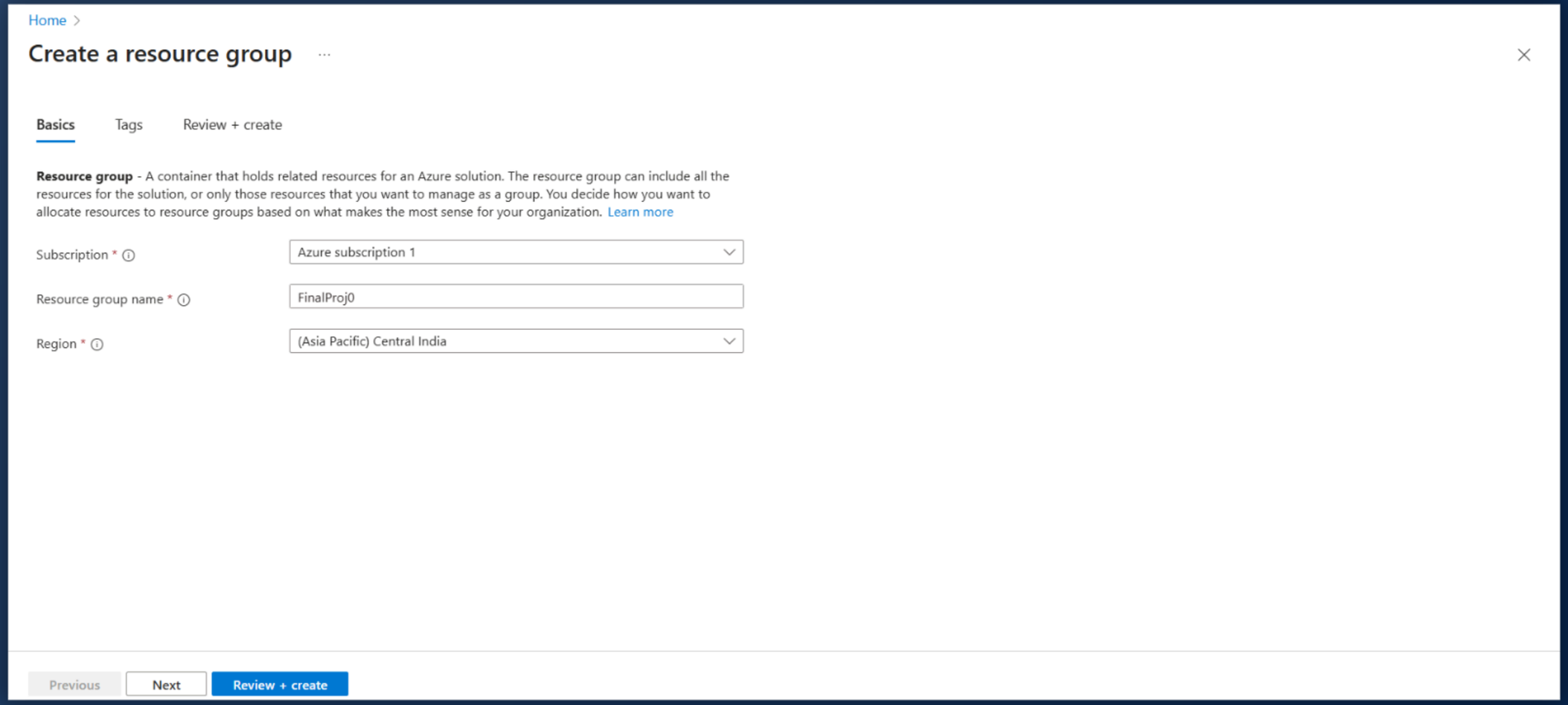
PROJECT WORKFLOW



KEY FEATURES

- Scalability: Handles large data volumes
- Automation: End-to-end pipeline automation
- Reliability: Delta Lake ensures ACID transactions
- Analytics-ready: High quality data for BI tools

RESOURCE GROUP CREATION



STORAGE ACCOUNT CREATION

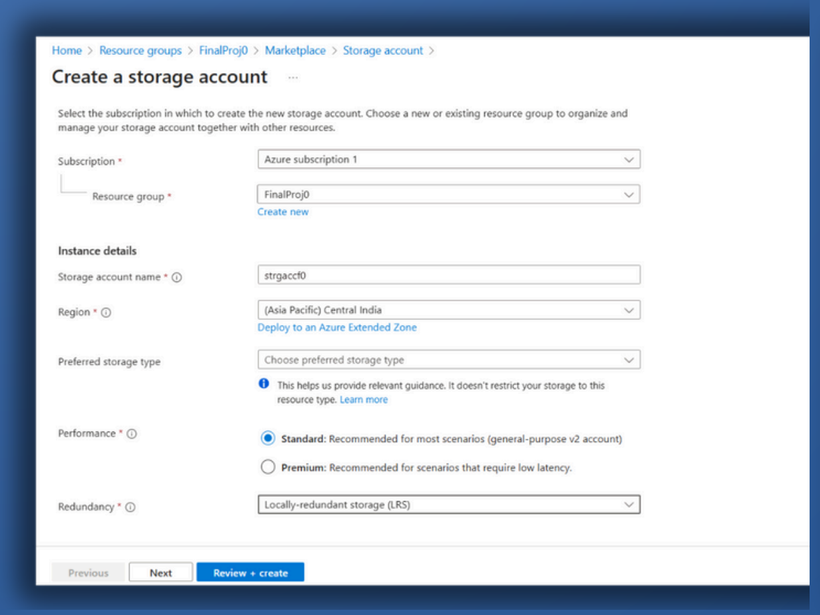


Fig1: Storage Account Creation

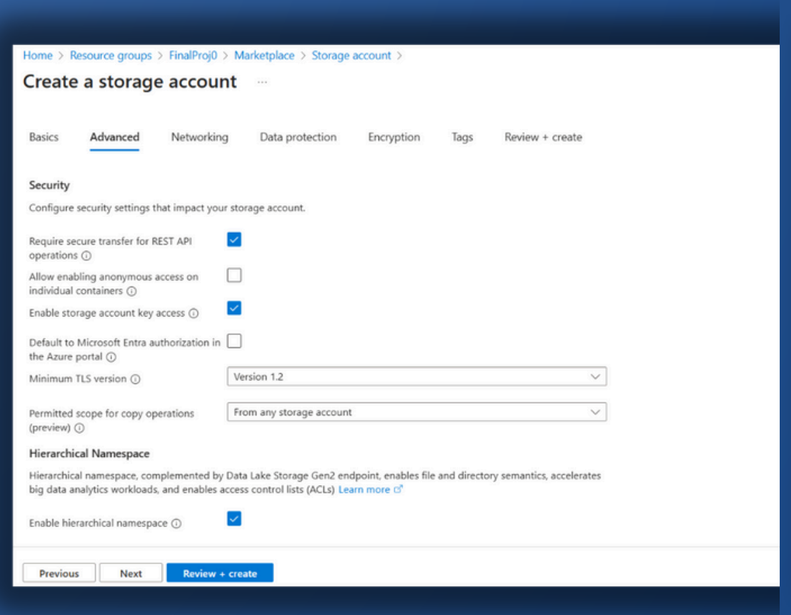


Fig2 : Storage Account Creation

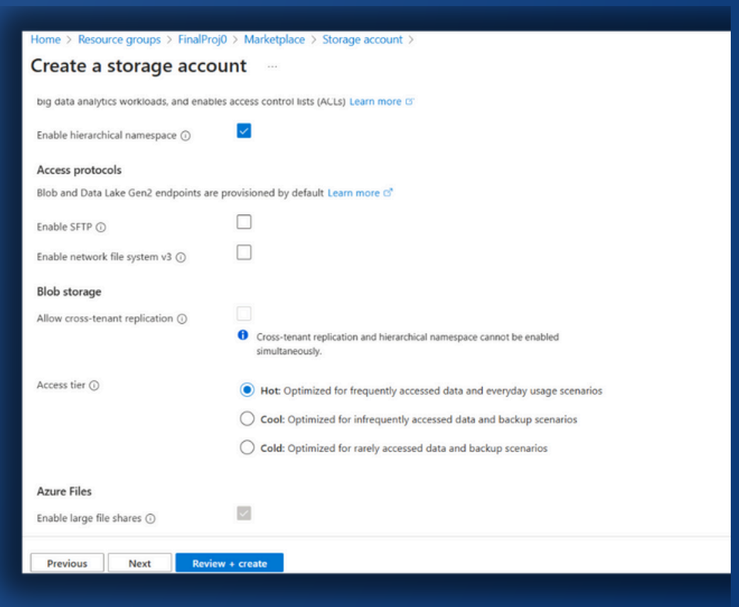
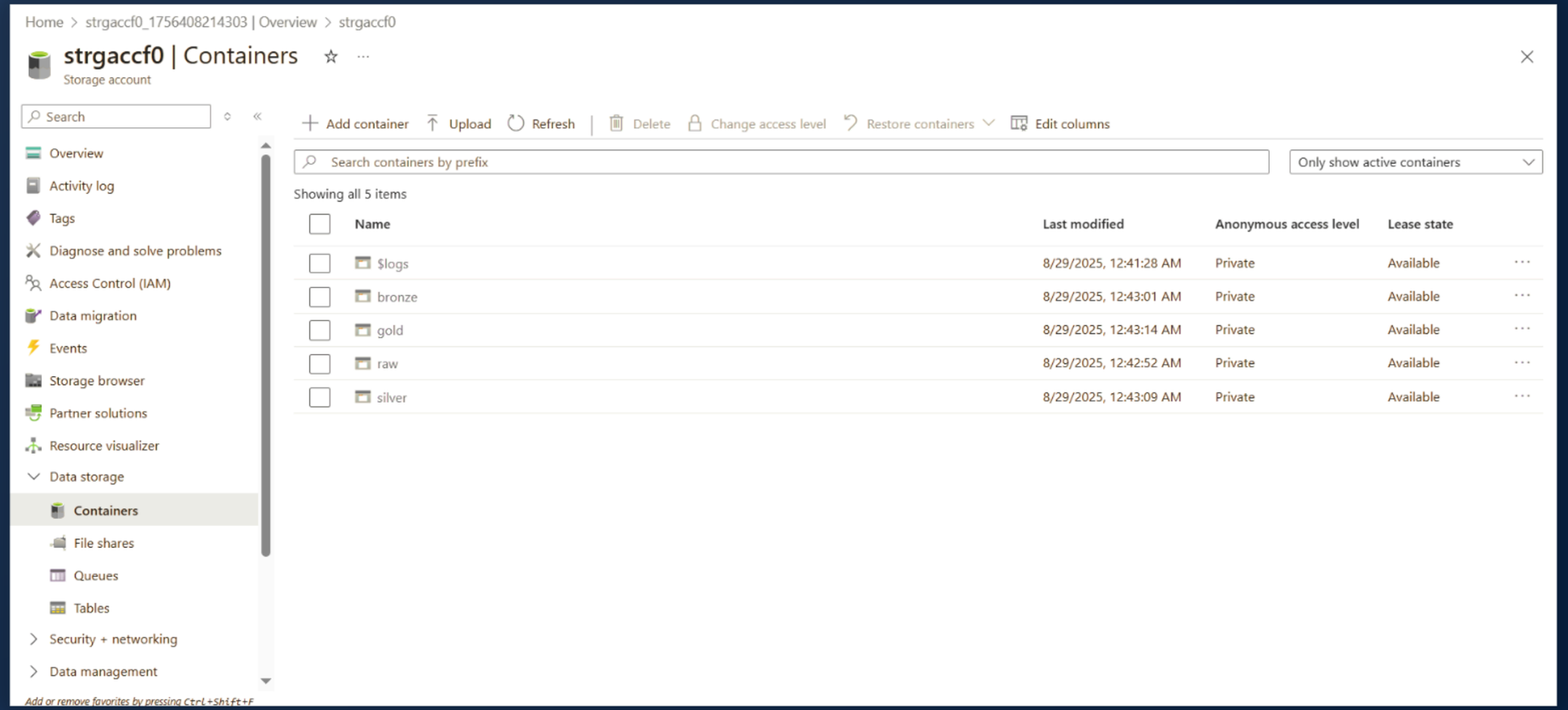


Fig3: Storage Account Creation

CREATION OF REQUIRED CONTAINERS



UPLOADING PARAMETER CONTAINED JSON FILE

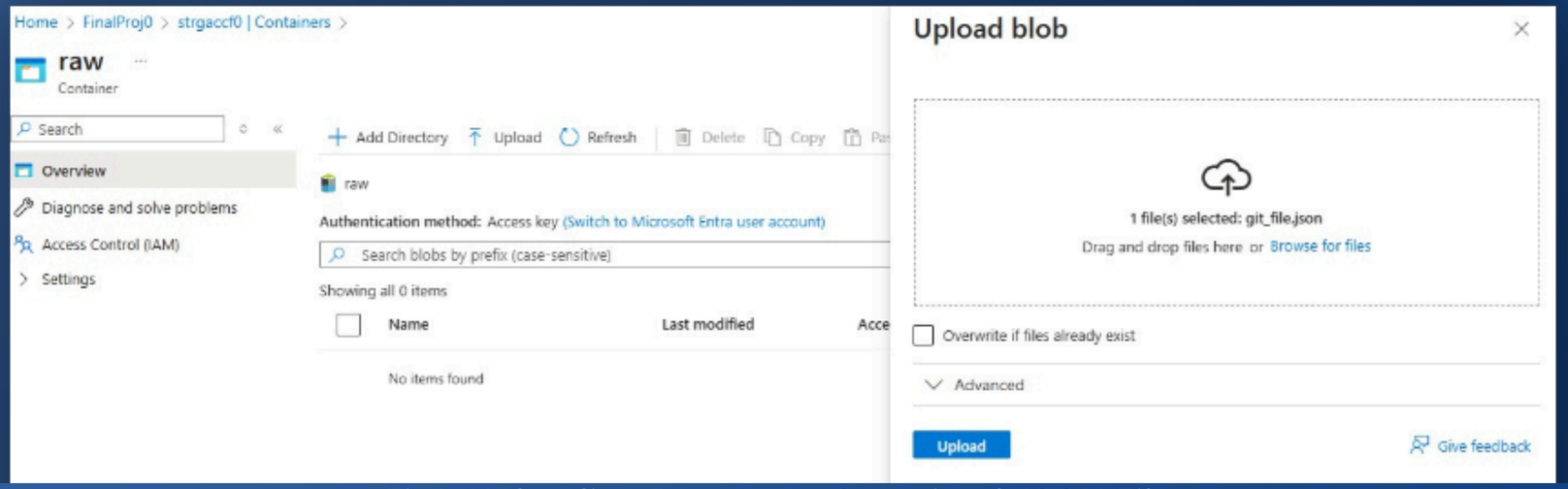


Fig1: Uploading Parameter Contained JSON File

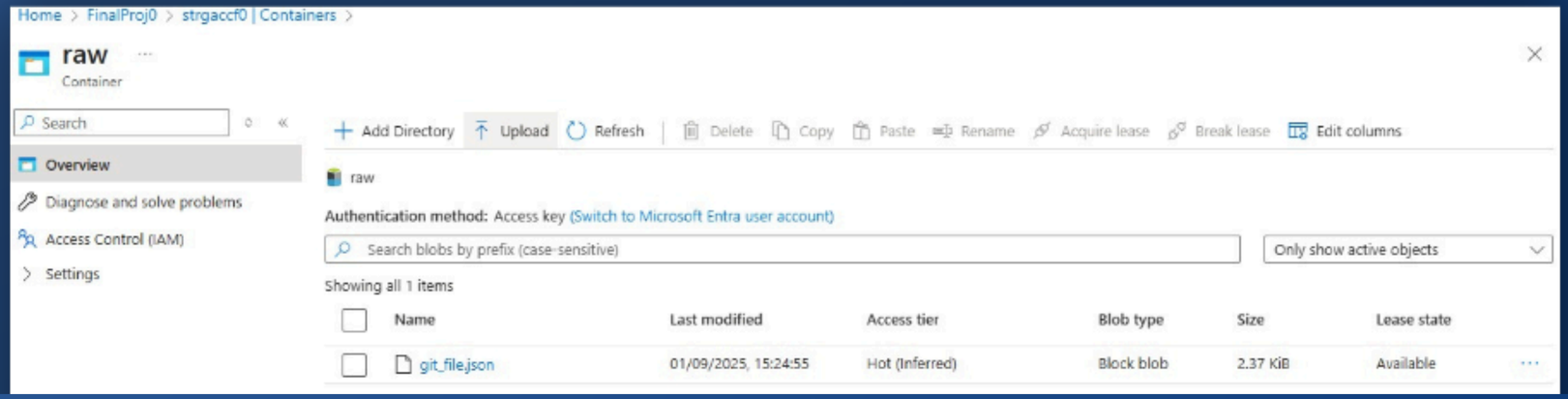


Fig2: Successfully Uploaded

CREATION OF AZURE DATA FACTORY

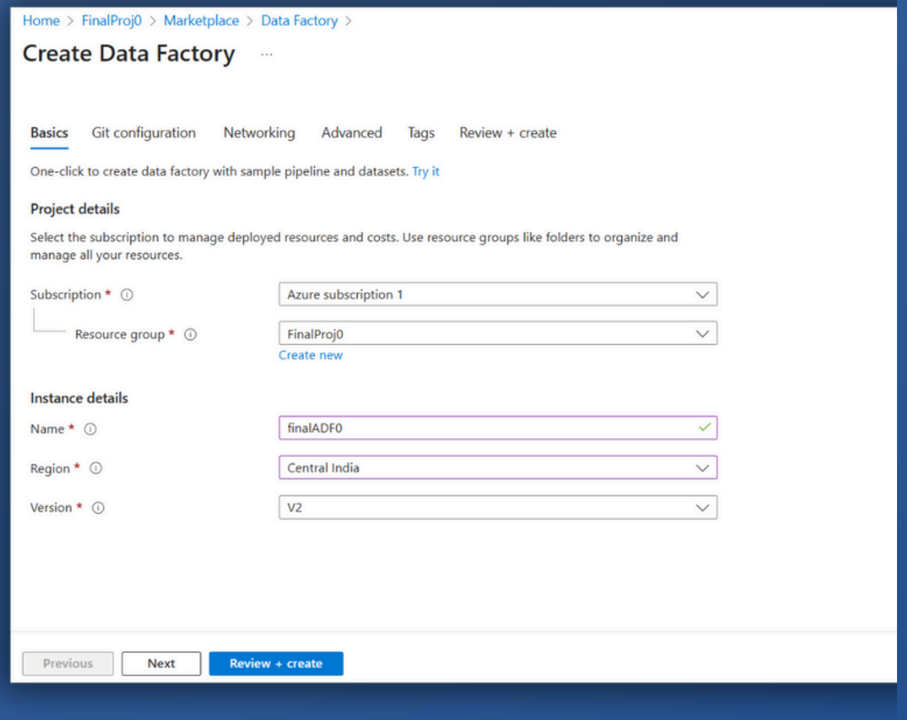


Fig: Creation of Azure Data Factory

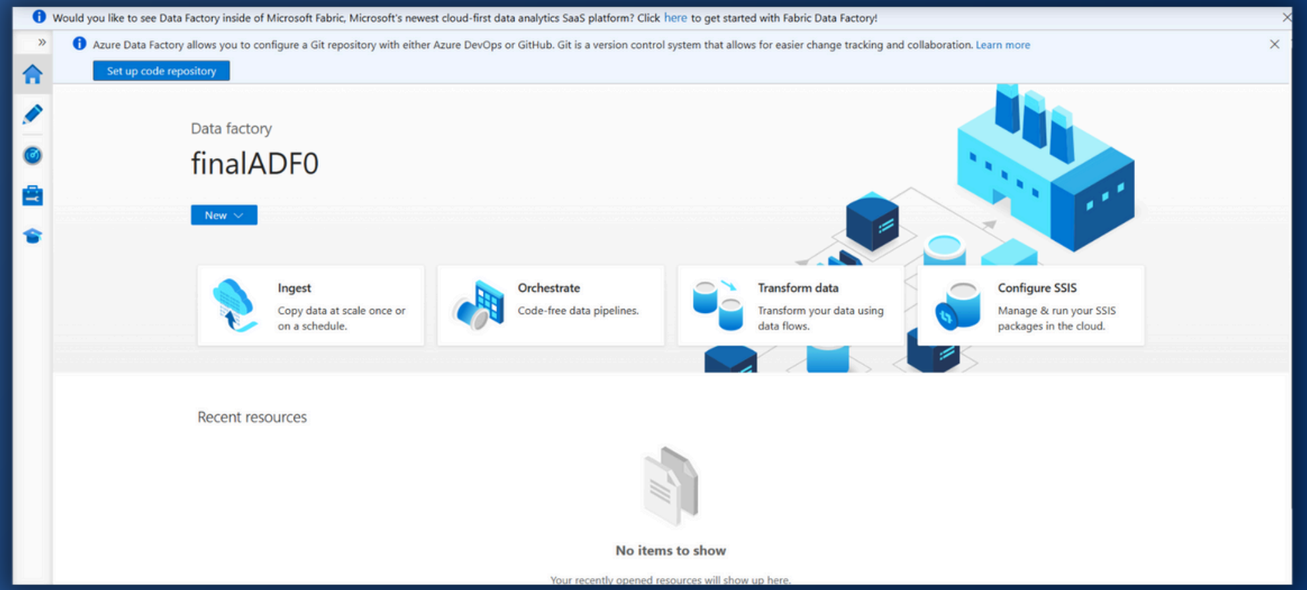


Fig: Successful launching of the ADF

DATA INGESTION PIPELINE

Lookup Activity and Output :

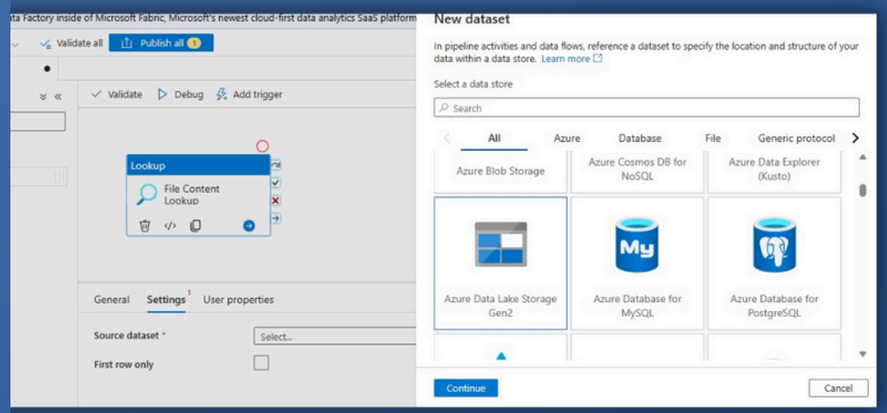


Fig: Dataset Creation for Lookup

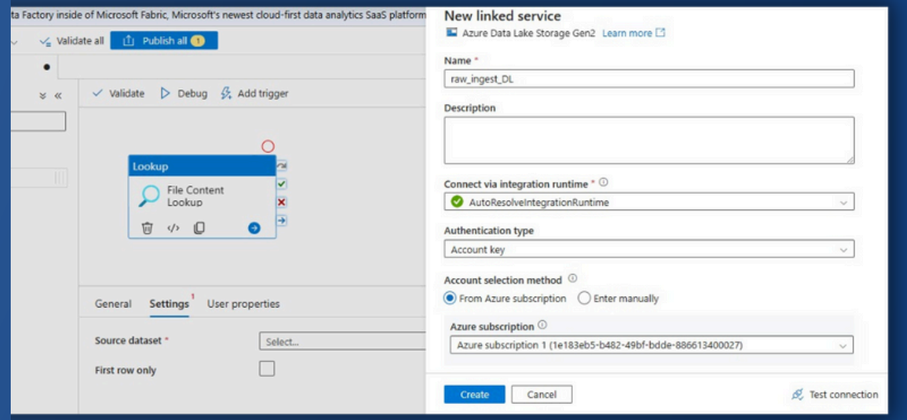


Fig: Linked Service Creation

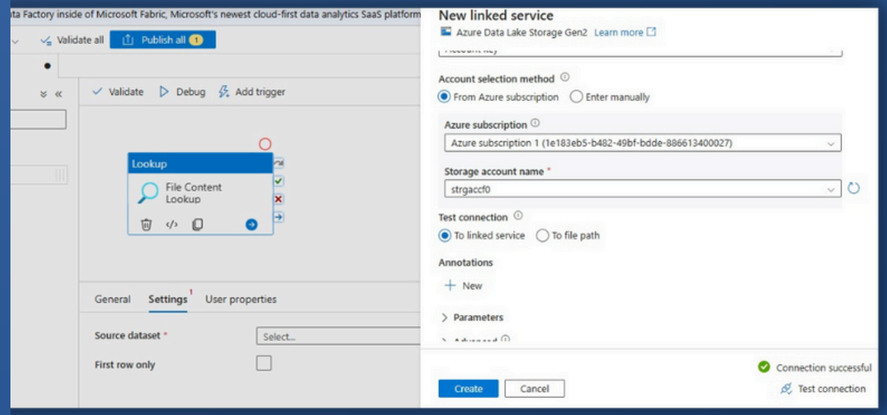


Fig: Linked Service Creation

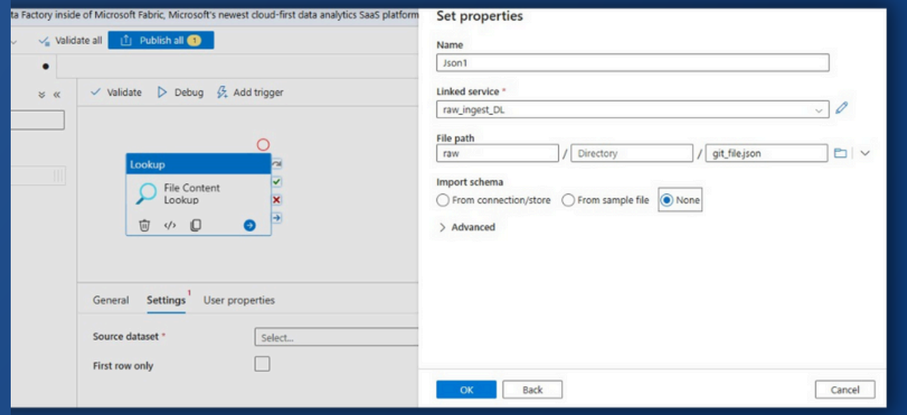


Fig: Selecting GitHub file in Lookup (raw container)

DATA INGESTION PIPELINE CONTINUED...

Lookup Activity and Output :

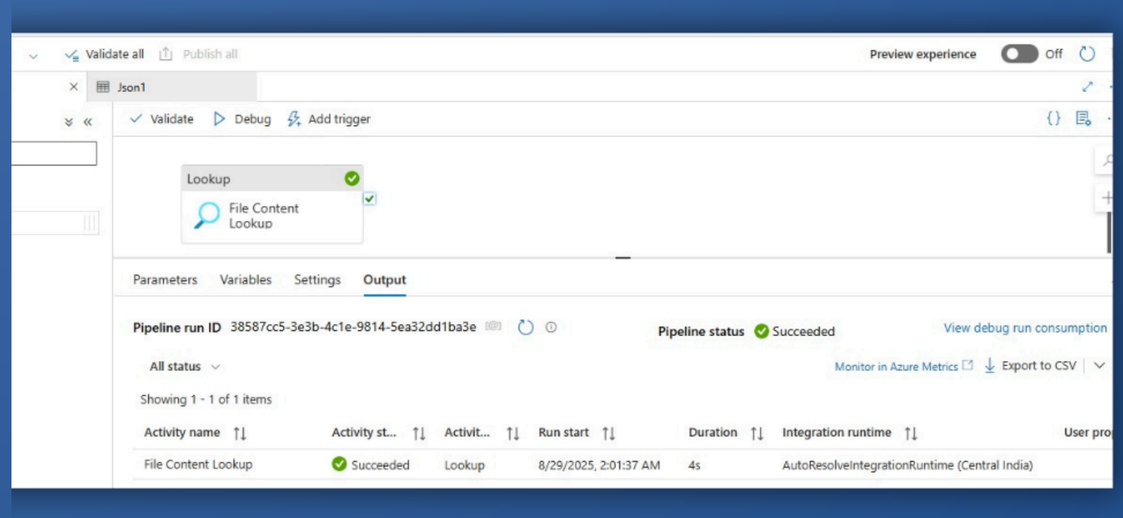


Fig: Lookup Activity run successfully

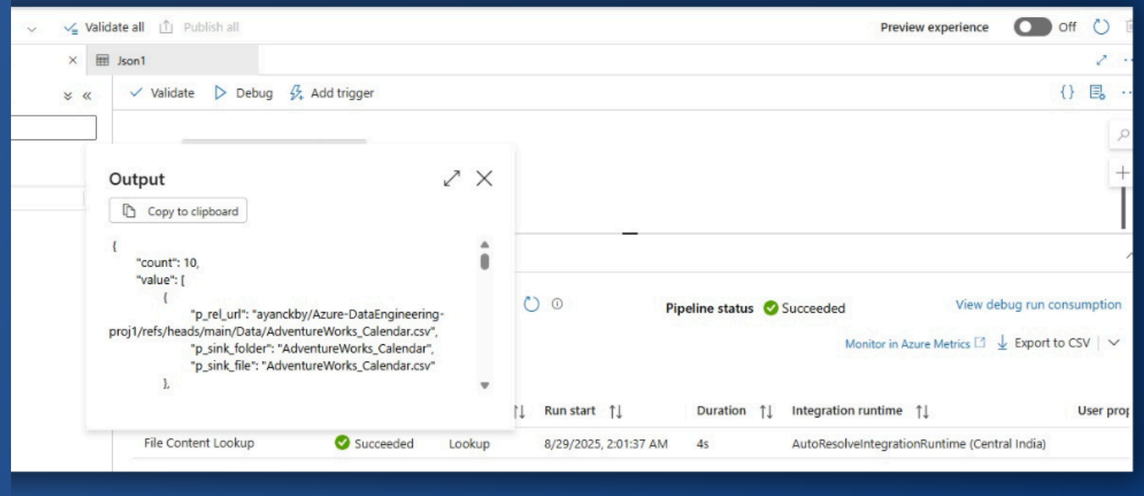


Fig: Lookup Output

DATA INGESTION PIPELINE CONTINUED...

ForEach Activity :

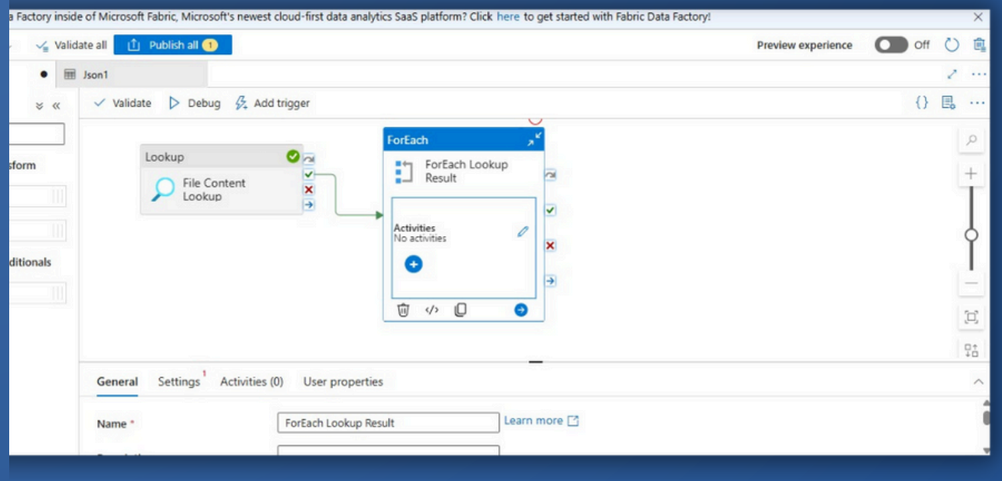


Fig: Connecting ForEach Activity with Lookup

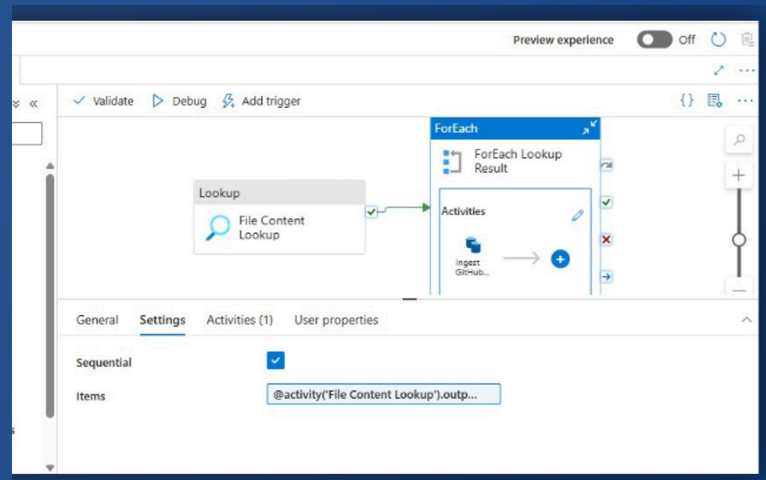


fig: Iterating over Lookup Output

DATA INGESTION PIPELINE CONTINUED...

Copy Data Activity:

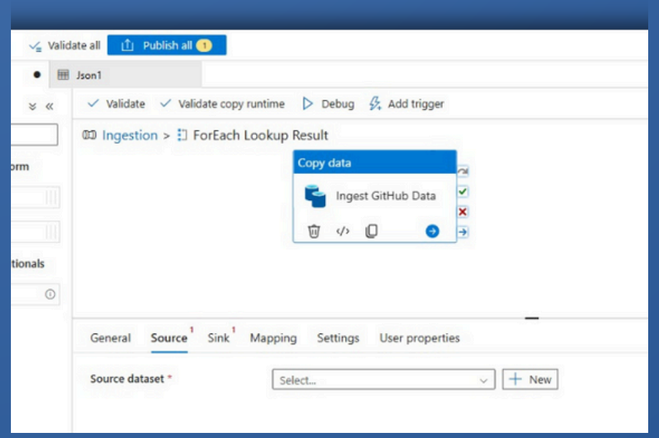


Fig: Selecting Copy data Activity and rename with ‘Ingest GitHub Data’

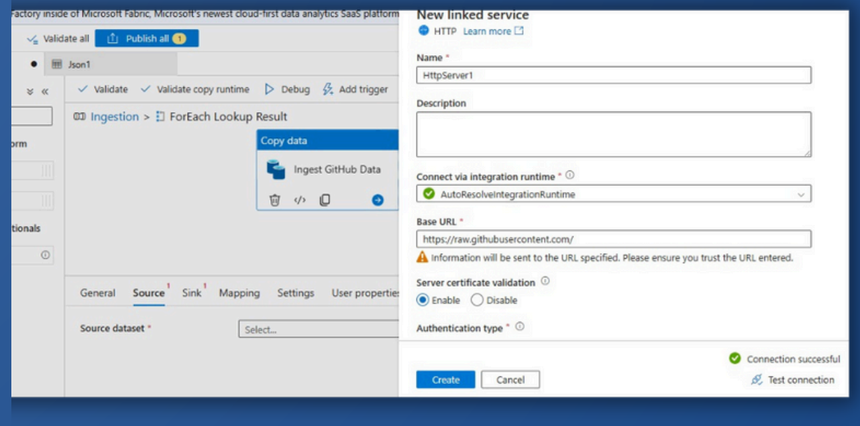


Fig: Linked Service Creation - GitHub as Source

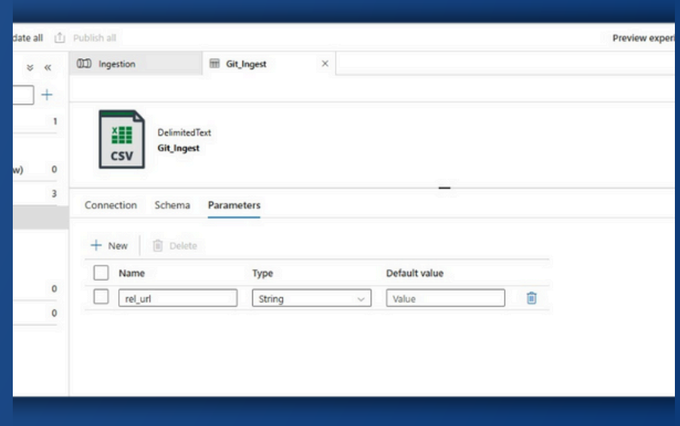


Fig: Parameter Creation for Source

DATA INGESTION PIPELINE CONTINUED...

Copy Data Activity Continued :

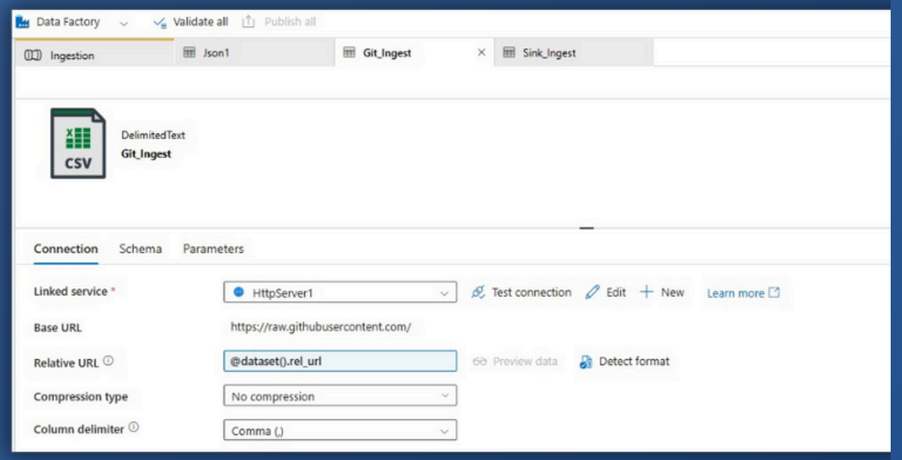


Fig: Relative URL Setup

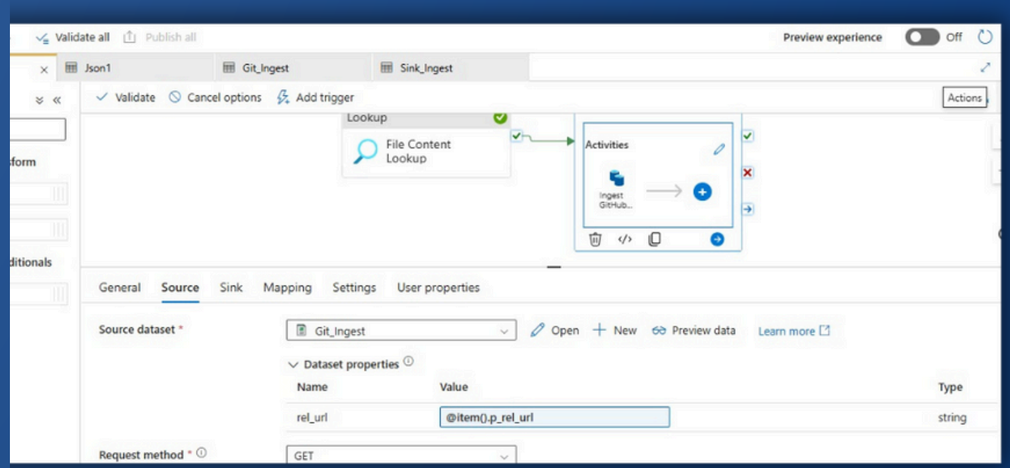


Fig: Assigning value to Relative URL Parameter

DATA INGESTION PIPELINE CONTINUED...

Copy Data Activity Continued :

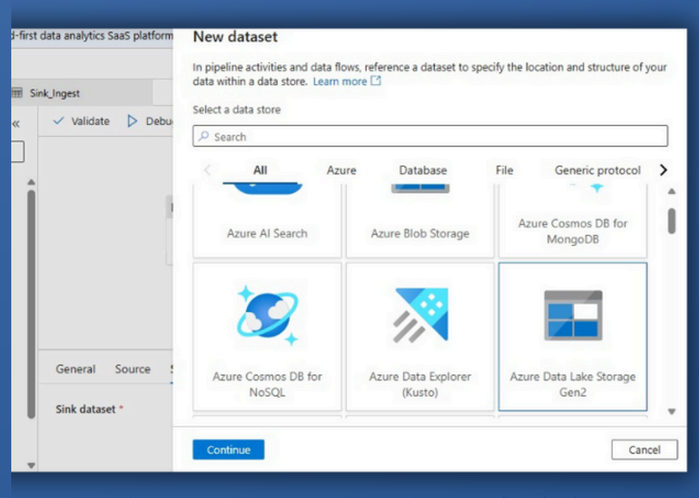


Fig: Dataset Creation of Sink

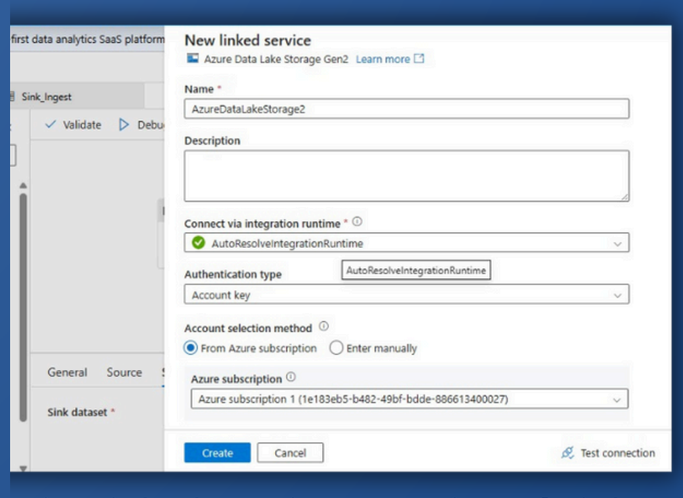


Fig: Linked Service Creation of Sink

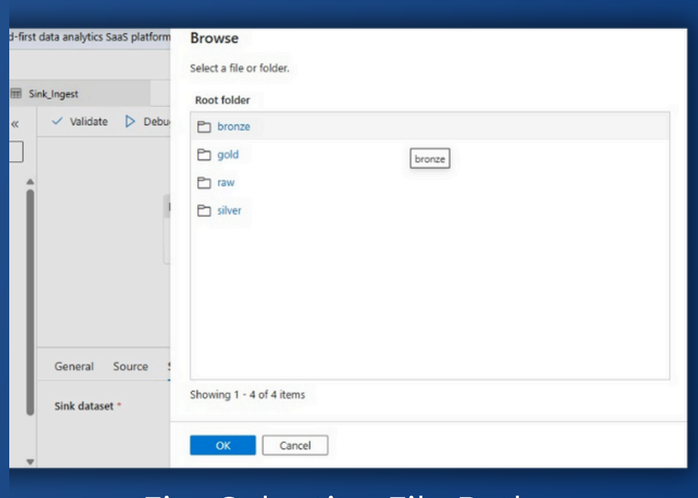


Fig: Selecting File Path

DATA INGESTION PIPELINE CONTINUED...

Copy Data Activity Continued :

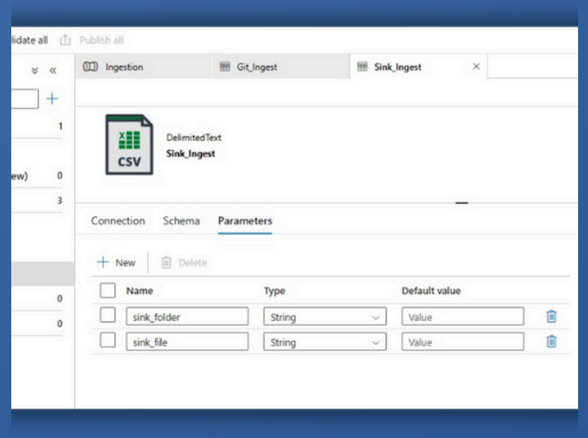


Fig: Creation of Parameters for Sink

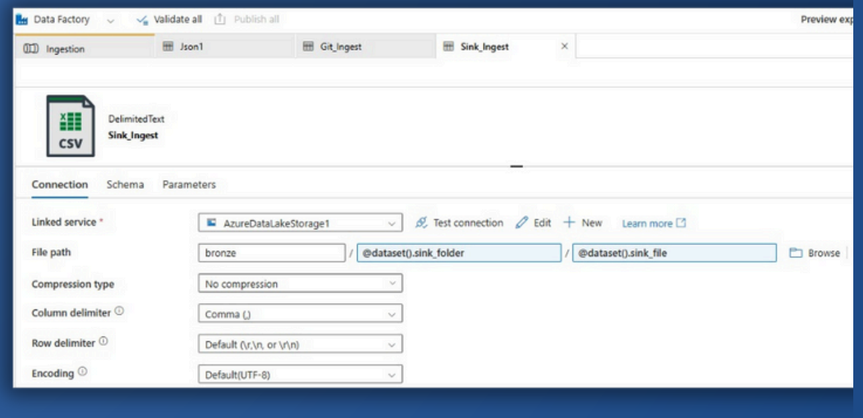


Fig: File path

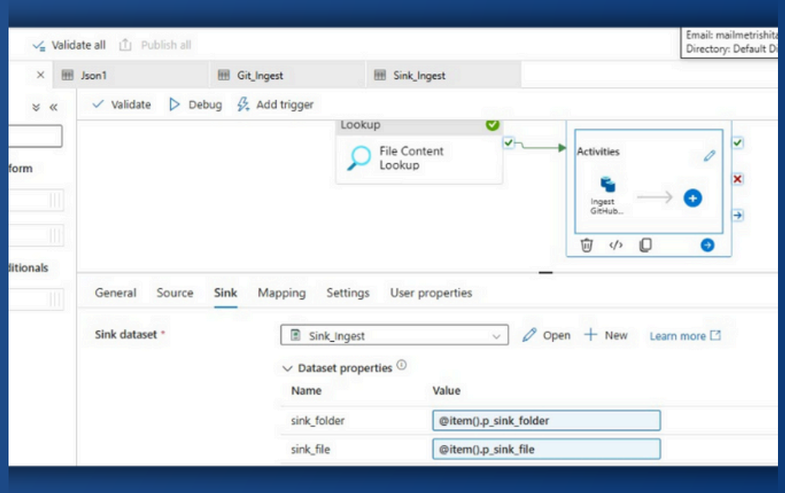


Fig: Assigning Value

DATA INGESTION PIPELINE CONTINUED...

Pipeline Output

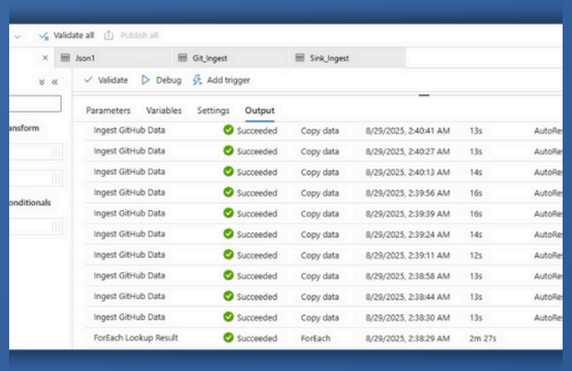


Fig: Pipeline Execution Successful

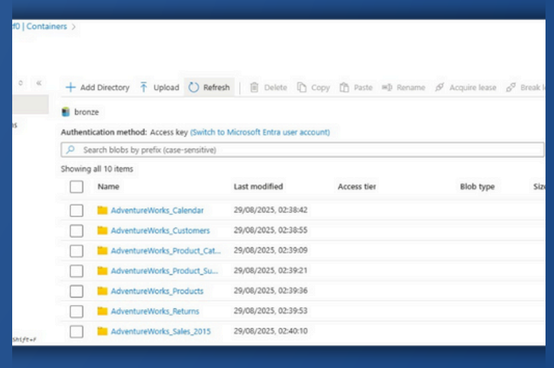


Fig: Successful Data Ingestion to Bronze Container

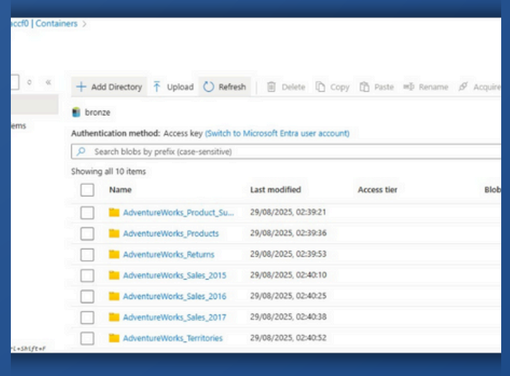
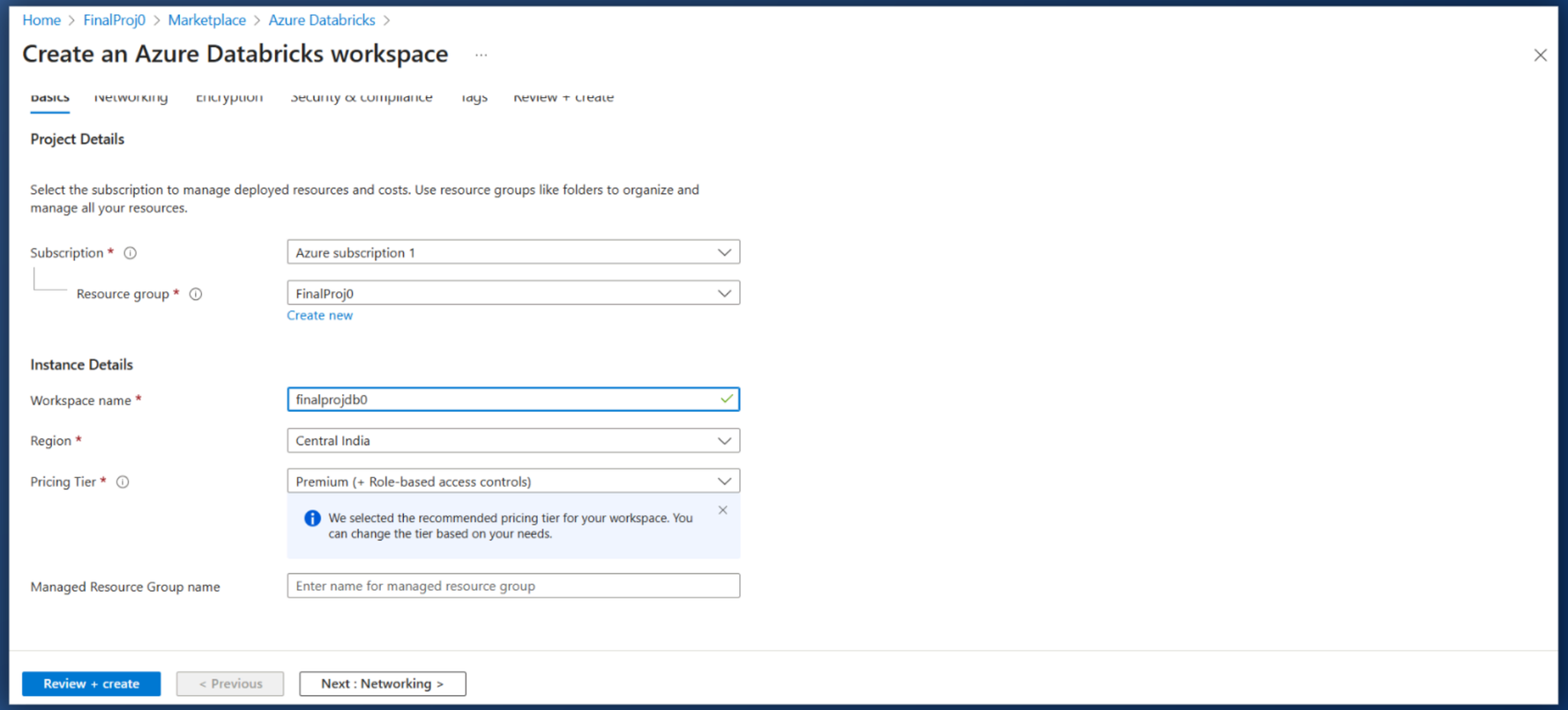


Fig: Successful Data Ingestion to Bronze Container

CREATION OF AZURE DATABRICKS



APP REGISTRATION AND ROLE ASSIGNMENT

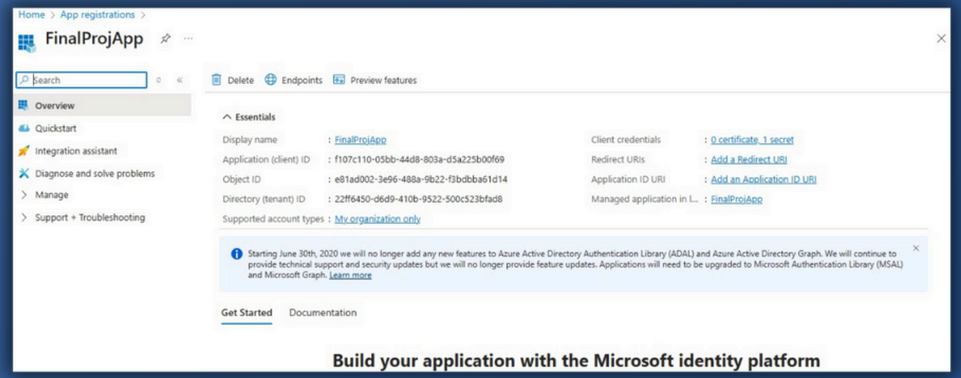


Fig: Creating App

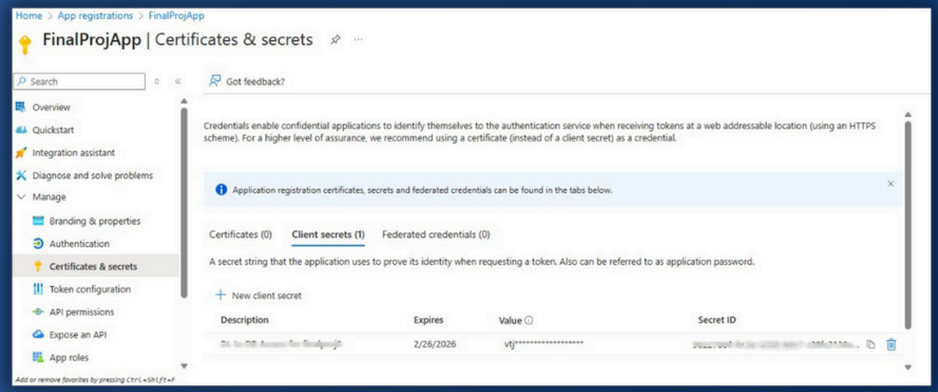


Fig: Providing Client Secret

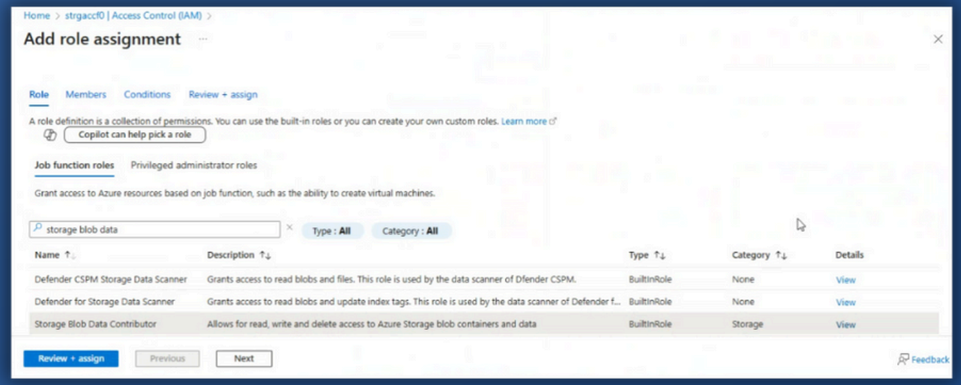


Fig: Role Assignment

Fig: Adding the App as Member

TRANSFORMATION USING DATABRICKS

Creating new Cluster:

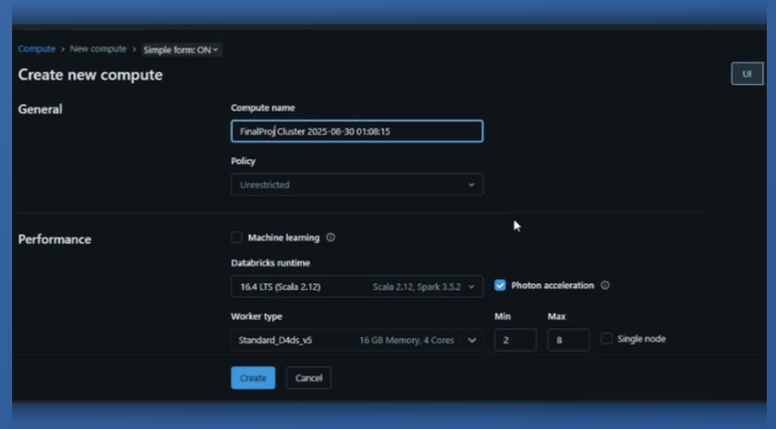


Fig: Creating New Cluster

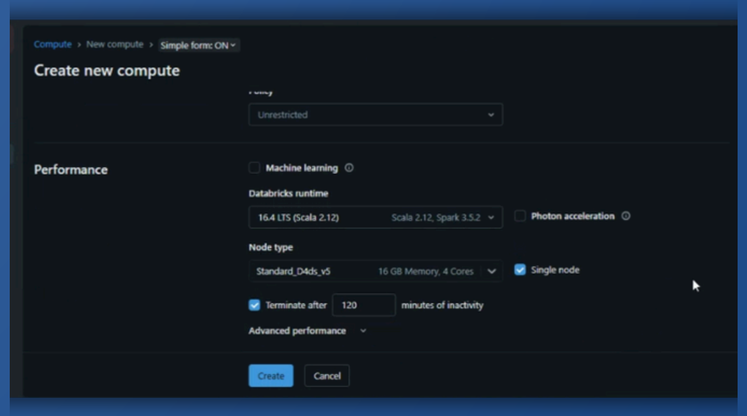


Fig: Creating New Cluster

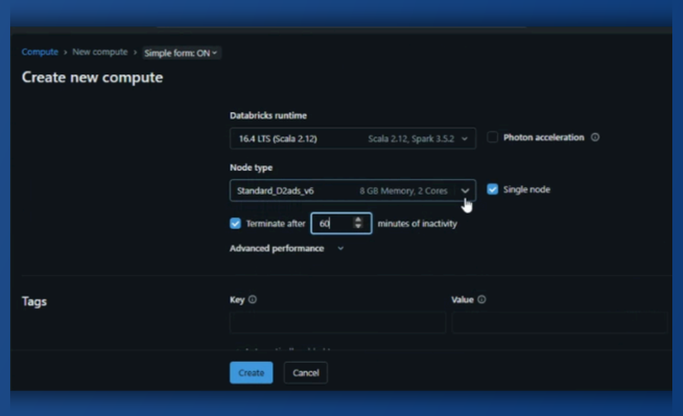


Fig: Creating New Cluster

TRANSFORMATION USING DATABRICKS CONTINUED...

Data Transformation and Load into Silver Layer

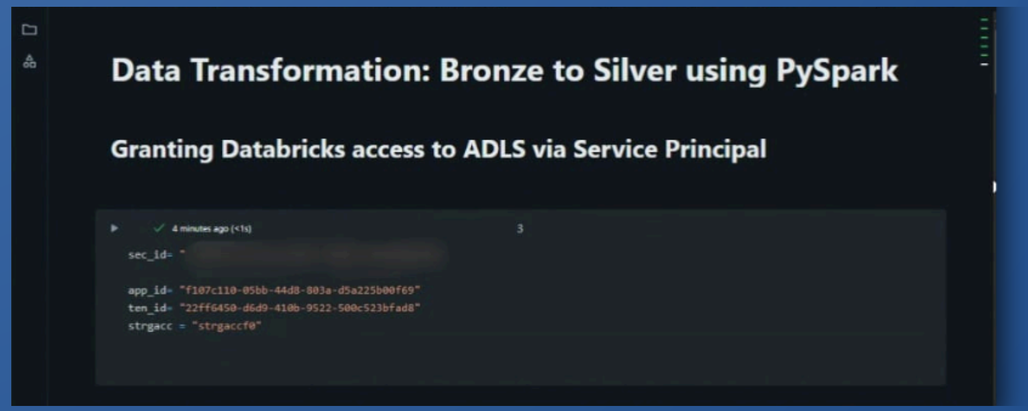


Fig:Granting DataBricks Access to ADLS via Service Principal

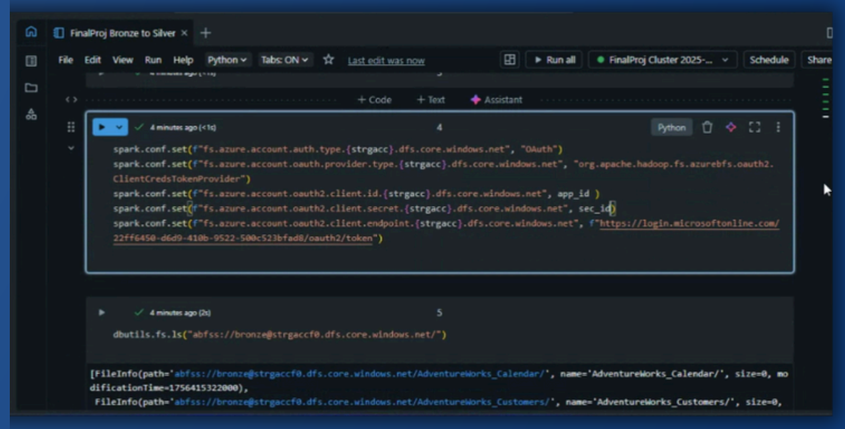


Fig:Granting DataBricks Access to ADLS via Service Principal

TRANSFORMATION USING DATABRICKS CONTINUED...

Data Transformation and Load into Silver Layer Continued...

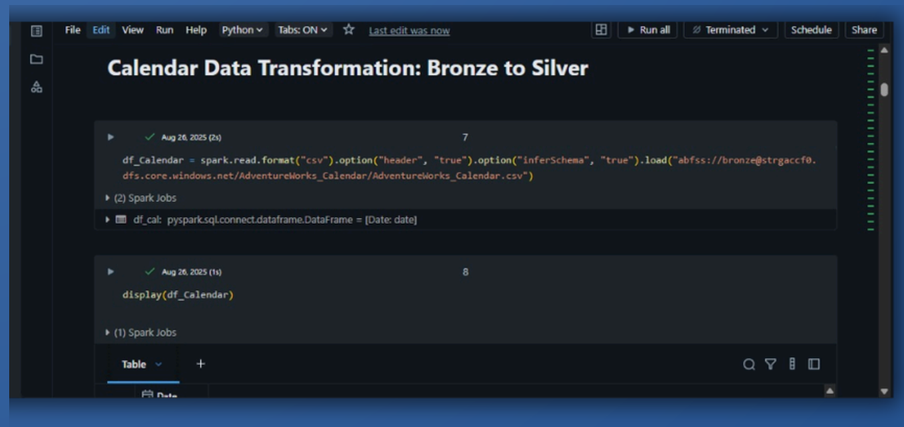


Fig: Calender Data Load Into DataBricks

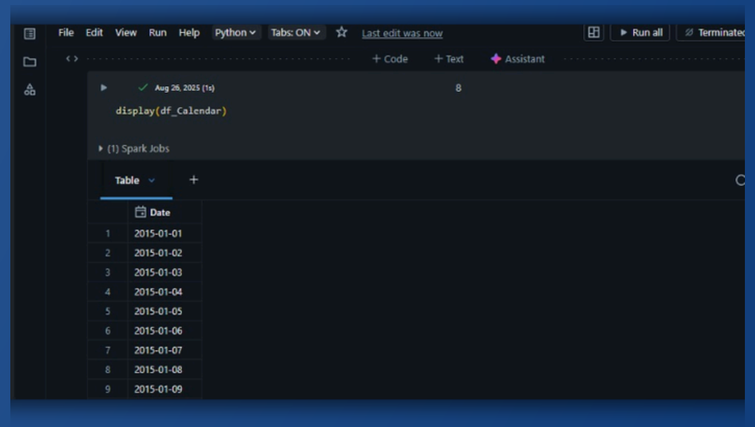


Fig: Calender Data Display

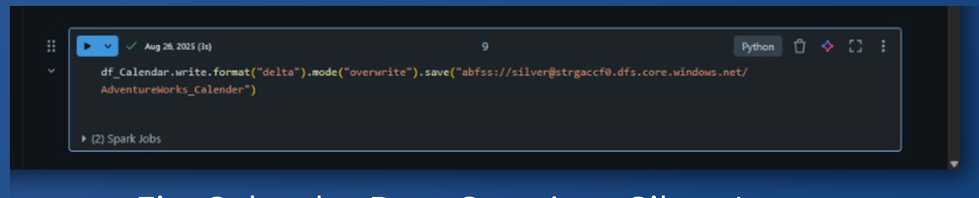


Fig: Calender Data Save Into Silver Layer

TRANSFORMATION USING DATABRICKS CONTINUED...

Data Transformation and Load into Silver Layer Continued...

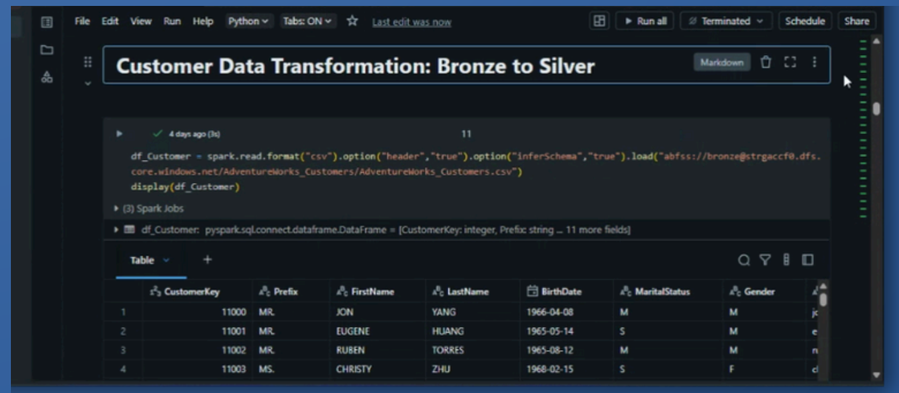


Fig: Customer Data Load Into DataBricks

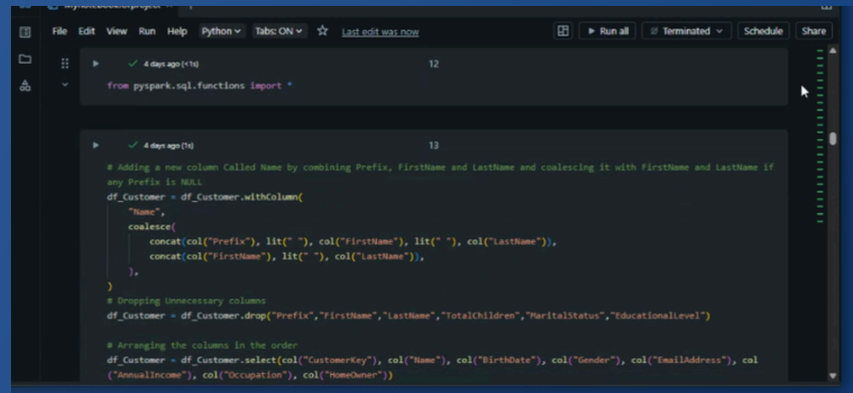


Fig: Customer Data Transformation

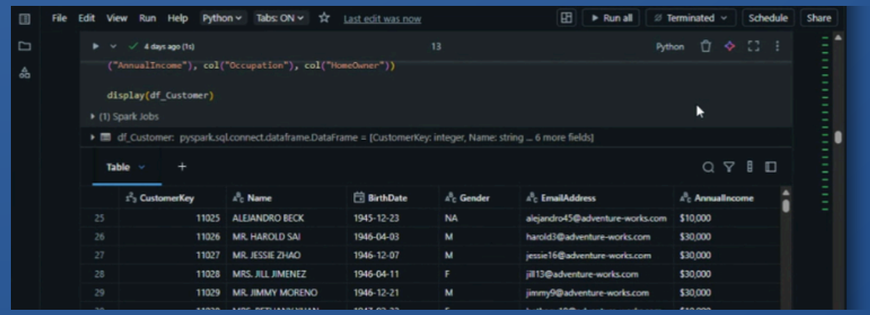


Fig: Customer Data Transformation

Fig: Customer Data Save Into Silver Layer

TRANSFORMATION USING DATABRICKS CONTINUED...

Data Transformation and Load into Silver Layer Continued...

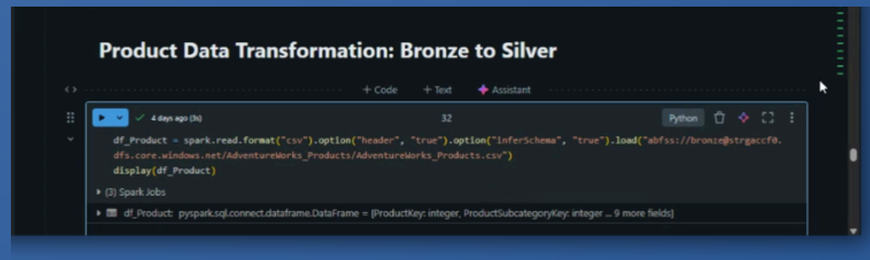


Fig: Product Data Load Into DataBricks

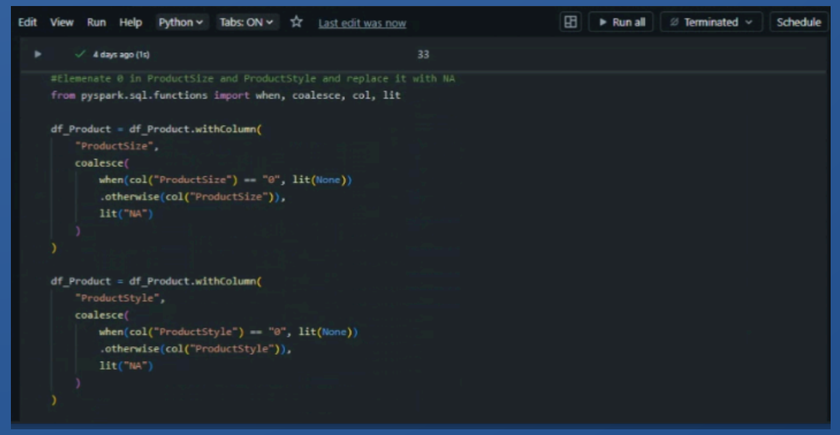


Fig: Product Data Transformation

	ProductKey	ProductSubcategoryKey	ProductSKU	ProductName	ModelName	ProductDe
1	214	31	HL-U509-R	Sport-100 Helmet, Red	Sport-100	Universal fit, w
2	215	31	HL-U509	Sport-100 Helmet, Black	Sport-100	Universal fit, w
3	218	23	SO-B909-M	Mountain Bike Socks, M	Mountain Bike Socks	Combination c
4	219	23	SO-B909-L	Mountain Bike Socks, L	Mountain Bike Socks	Combination c
5	220	31	HL-U509-B	Sport-100 Helmet, Blue	Sport-100	Universal fit, w
6	223	19	CA-1098	AWC Logo Cap	Cycling Cap	Traditional styl
7	226	21	LJ-0192-S	Long-Sleeve Logo Jersey, S	Long-Sleeve Logo Jers...	Unisex long-sl
8	229	21	LJ-0192-M	Long-Sleeve Logo Jersey, M	Long-Sleeve Logo Jers...	Unisex long-sl
9	232	21	LJ-0192-L	Long-Sleeve Logo Jersey, L	Long-Sleeve Logo Jers...	Unisex long-sl
10	235	21	LJ-0192-X	Long-Sleeve Logo Jersey, XL	Long-Sleeve Logo Jers...	Unisex long-sl
11	238	14	FR-R92R-62	HL Road Frame - Red, 62	HL Road Frame	Our lightest
12	241	14	FR-R92R-44	HL Road Frame - Red, 44	HL Road Frame	Our lightest
13	244	14	FR-R92R-48	HL Road Frame - Red, 48	HL Road Frame	Our lightest
14	247	14	FR-R92R-52	HL Road Frame - Red, 52	HL Road Frame	Our lightest
15						

Fig: Product Data Display Before Transformation

TRANSFORMATION USING DATABRICKS CONTINUED...

Data Transformation and Load into Silver Layer Continued...

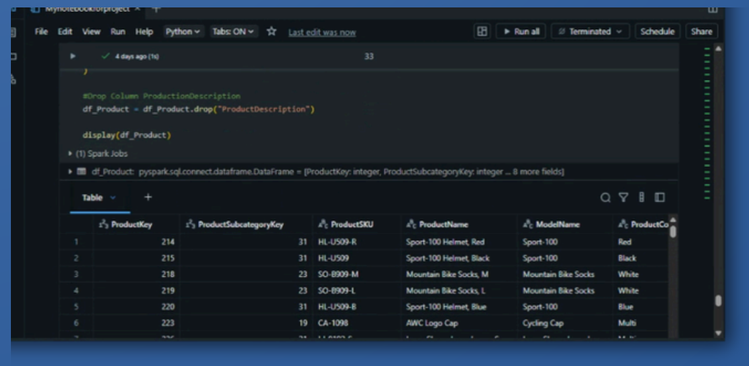


Fig: Product Data Display After Transformation

Fig: Product Data Save Into Silver Layer

TRANSFORMATION USING DATABRICKS CONTINUED...

Data Transformation and Load into Silver Layer Continued...

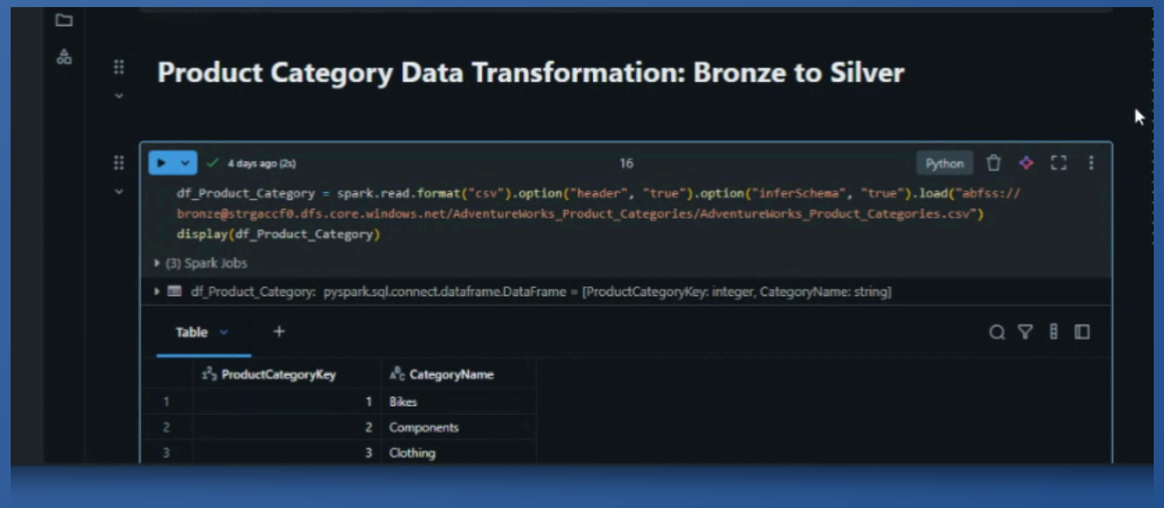


Fig: Product Catagory Data Load Into DataBricks and Display

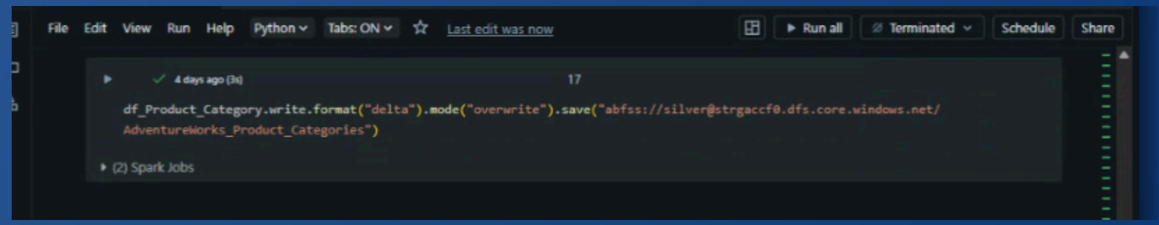


Fig: Product Catagory Data Save Into Silver Layer

TRANSFORMATION USING DATABRICKS CONTINUED...

Data Transformation and Load into Silver Layer Continued...

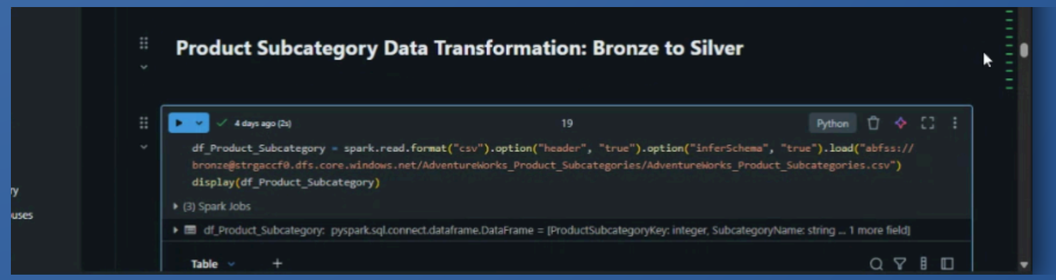


Fig: Product SubCatagory Data Load Into DataBricks

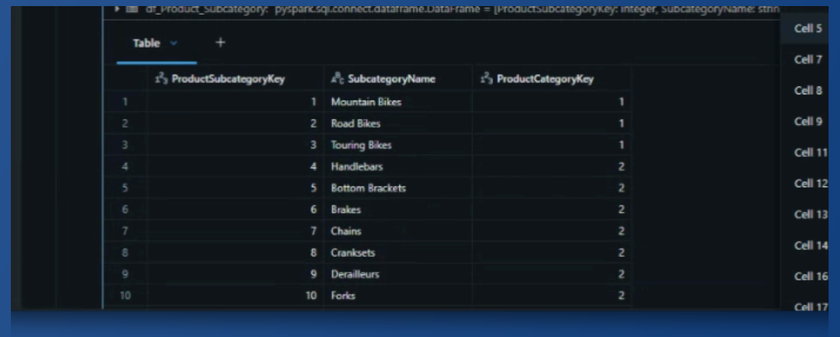


Fig: Product SubCatagory Data Display

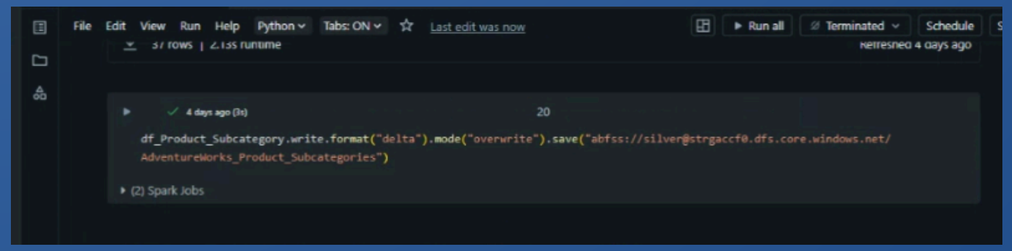


Fig: Product SubCatagory Data Save Into Silver Layer

TRANSFORMATION USING DATABRICKS CONTINUED...

Data Transformation and Load into Silver Layer Continued...

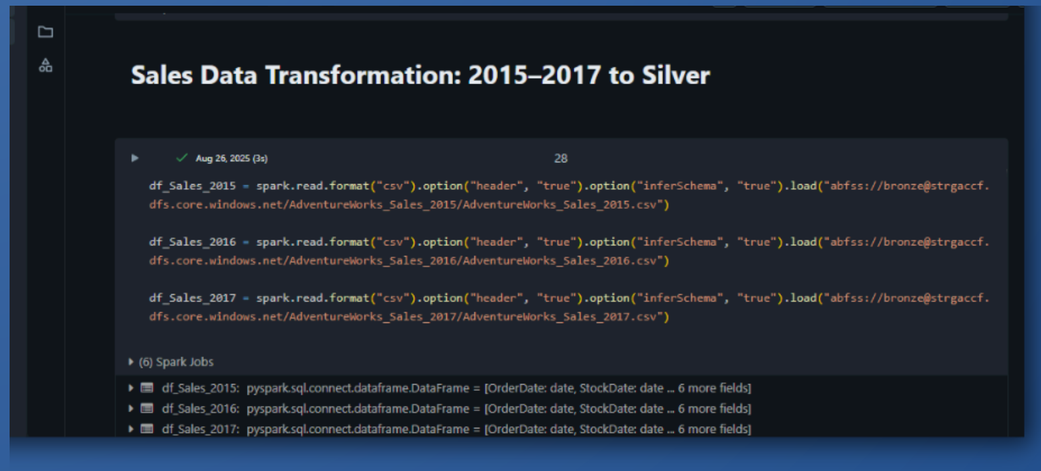


Fig: Sales Data(2015-17) Load Into DataBricks

Fig: Union Operation of Sales Data 2015,2016 and 2017

Fig: After Union Operation Save Into Silver Layer

TRANSFORMATION USING DATABRICKS CONTINUED...

Data Transformation and Load into Silver Layer Continued...

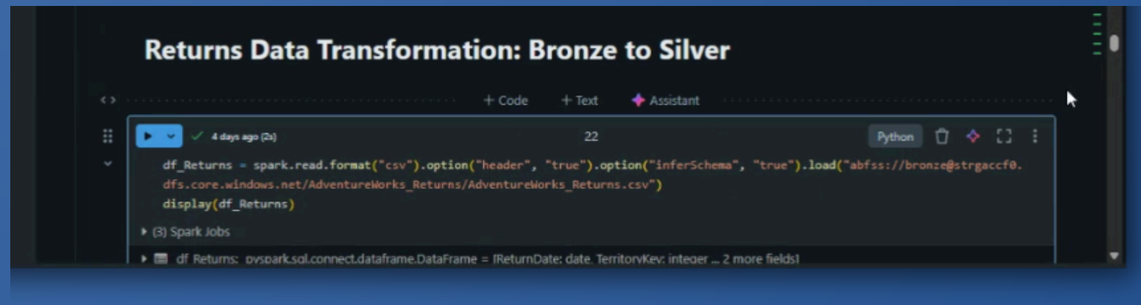


Fig: Returns Data Load Into DataBricks

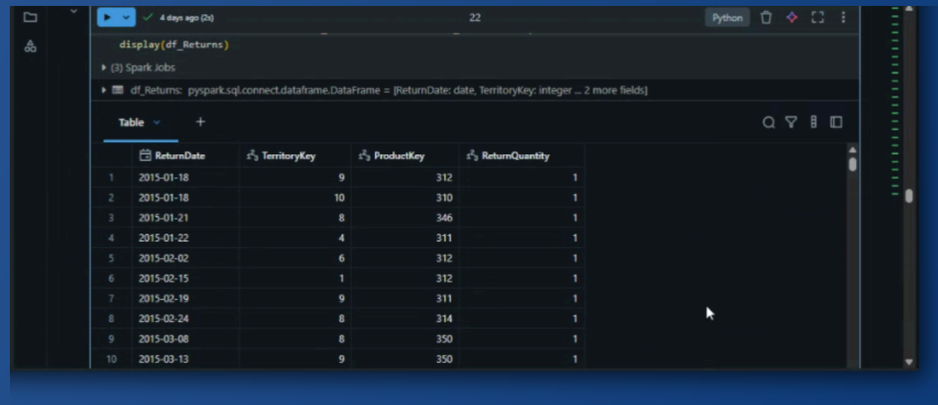


Fig: Returns Data Display

Fig: Returns Data Save Into Silver Layer

TRANSFORMATION USING DATABRICKS CONTINUED...

Data Transformation and Load into Silver Layer Continued...

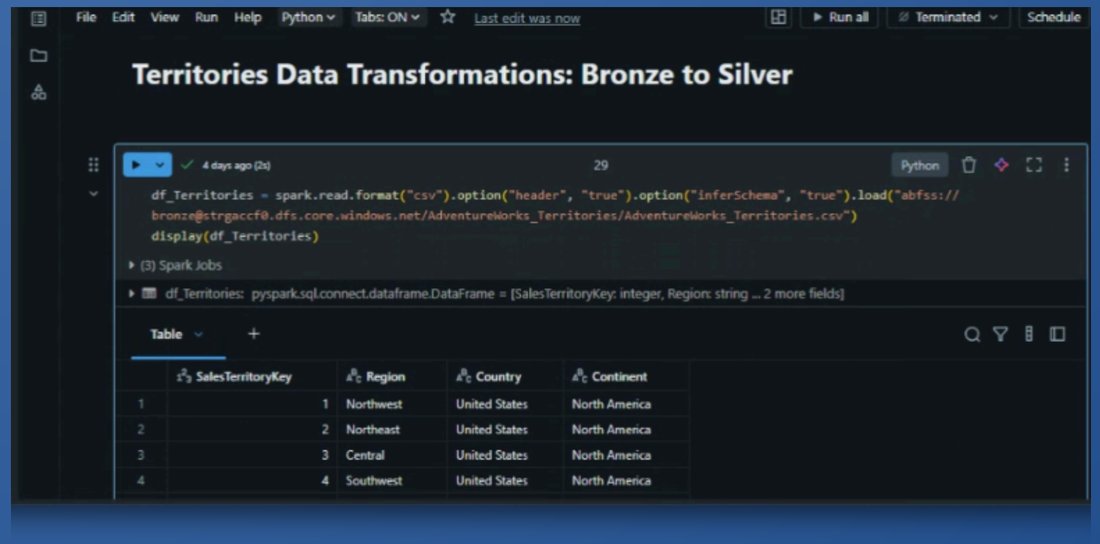


Fig: Territories Data Load Into DataBricks

Fig: Territories Data Save Into Silver Layer

TRANSFORMATION USING DATABRICKS CONTINUED...

Successful Data Ingestion: Bronze to Silver

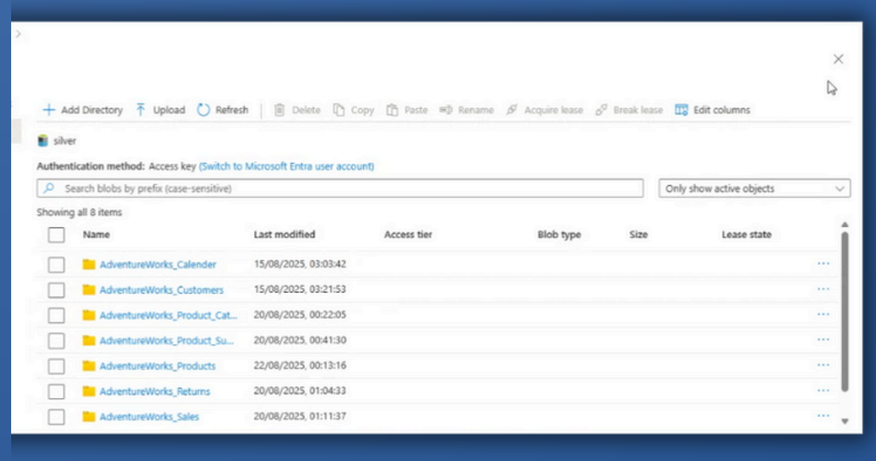


Fig: Transformed Data in Silver Layer

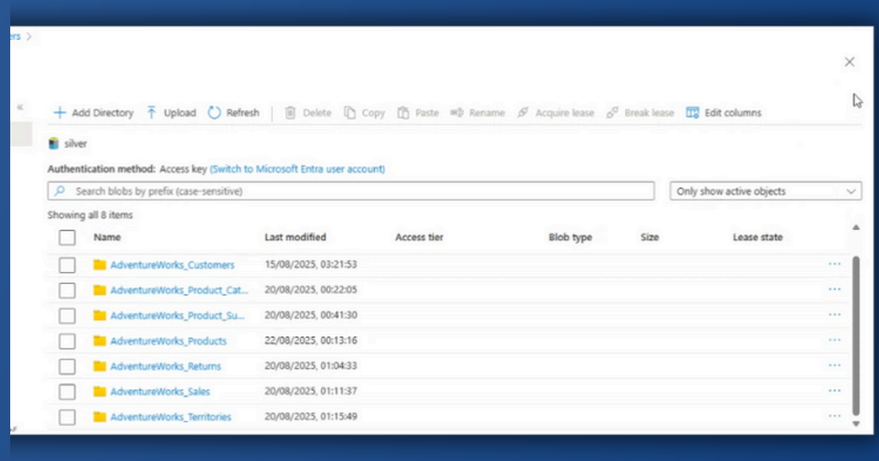


Fig: Transformed Data in Silver Layer

TRANSFORMATION USING DATABRICKS CONTINUED...

Data Aggregation and Storing into Gold Layer:

Fig: Importing All Functions

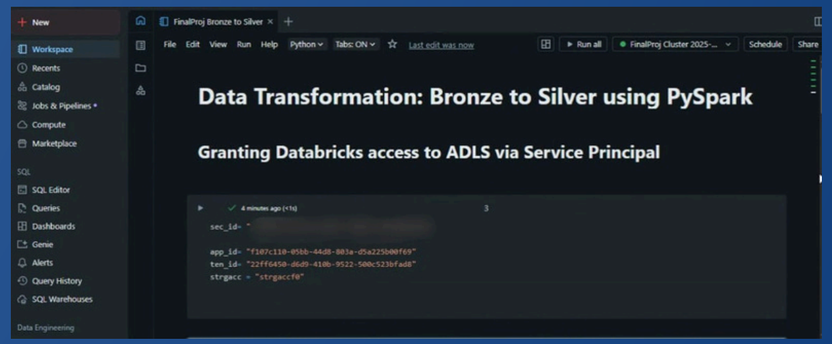


Fig: Granting DataBricks Access to ADLS via Service Principal

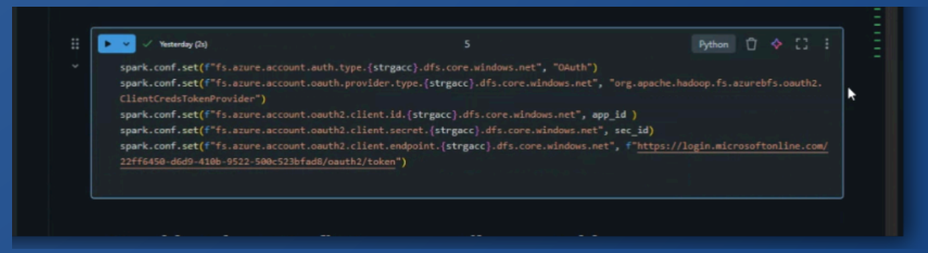


Fig: Granting DataBricks Access to ADLS via Service Principal

TRANSFORMATION USING DATABRICKS CONTINUED...

Data Aggregation and Storing into Gold Layer Continued...

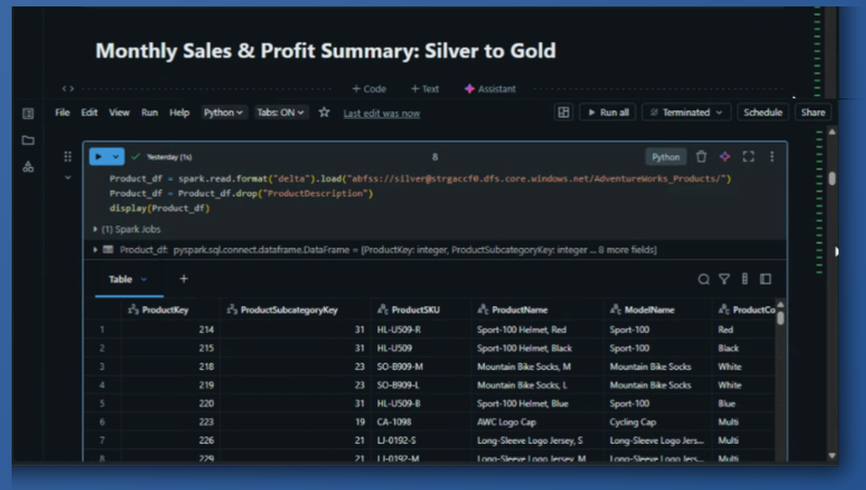


Fig: Product Data Load Into DataBricks From Silver Layer

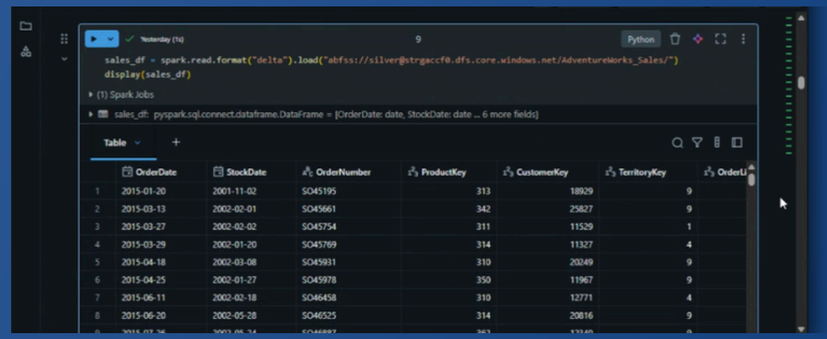


Fig: Sales Data Load Into DataBricks From Silver Layer

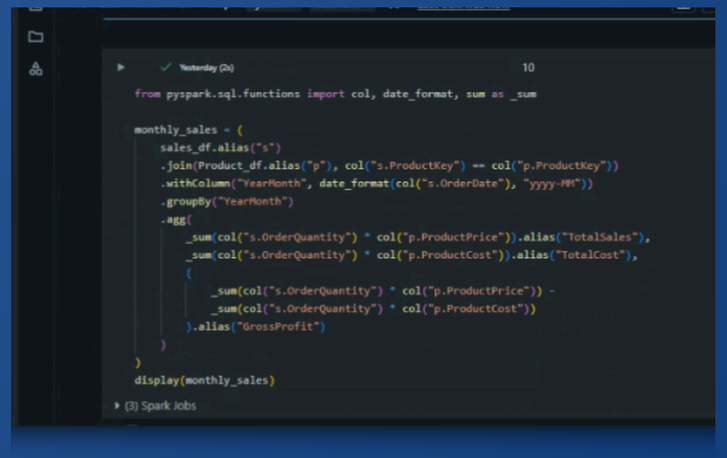


Fig: Monthly Sales Aggregation

TRANSFORMATION USING DATABRICKS CONTINUED...

Data Aggregation and Storing into Gold Layer Continued...

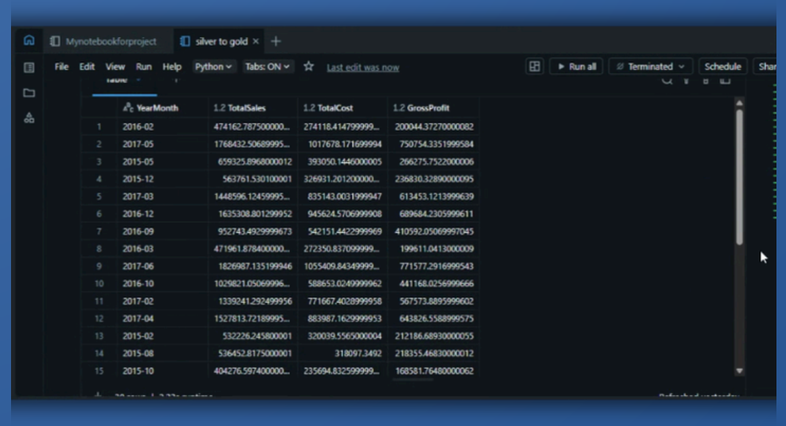


Fig: Monthly Sales Data Display

Fig: Monthly Sales Data Save Into Gold Layer

TRANSFORMATION USING DATABRICKS CONTINUED...

Data Aggregation and Storing into Gold Layer Continued...:

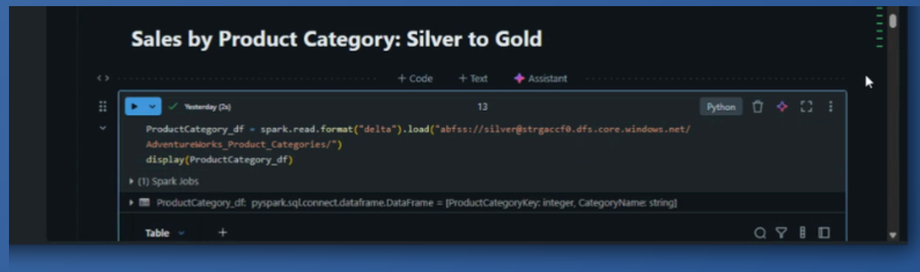


Fig: Product Category Data Load Into DataBricks From Silver Layer

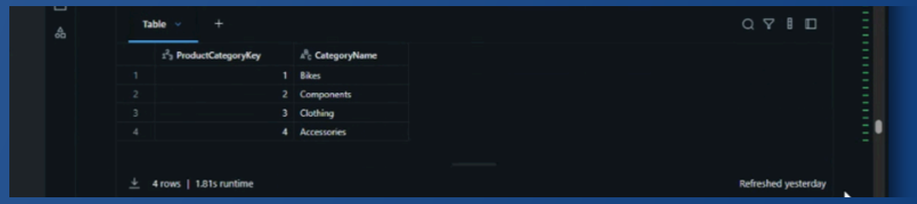


Fig: Product Category Data Display

Fig: Product SubCategory Data Load Into DataBricks From Silver Layer

Fig: Product SubCategory Data Display

TRANSFORMATION USING DATABRICKS CONTINUED...

Data Aggregation and Storing into Gold Layer Continued...:

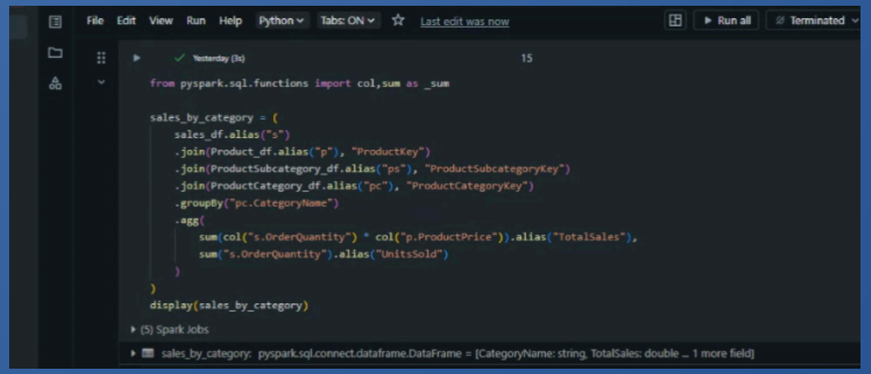


Fig: Sales By Product Category Aggregation

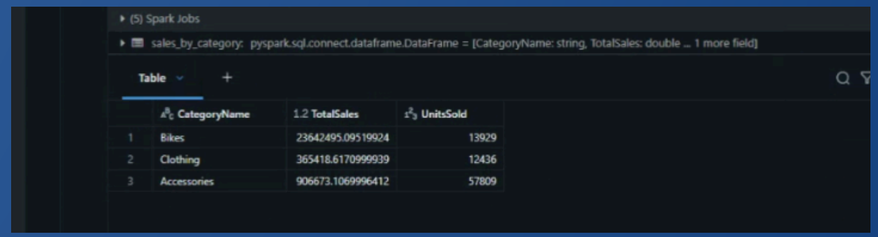


Fig: Sales By Product Category Data Display

Fig: Sales By Product Category Data Save Into Gold Layer

TRANSFORMATION USING DATABRICKS CONTINUED...

Data Aggregation and Storing into Gold Layer Continued...

Customer Insights (Lifetime Value): Silver to Gold

```
4 days ago (3s)                                                    18
from pyspark.sql import functions as F

customer_sales = (
    sales_df.alias("s")
    .join(Product_df.alias("p"), "ProductKey")
    .groupBy("s.CustomerKey")
    .agg(
        F.sum(F.col("s.OrderQuantity") * F.col("p.ProductPrice")).alias("TotalSpend"),
        F.countDistinct("s.OrderNumber").alias("TotalOrders"),
        F.avg(F.col("s.OrderQuantity") * F.col("p.ProductPrice")).alias("AvgOrderValue")
    )
)
display(customer_sales)
```

Fig: Customer Insights(Lifetime Value) Aggregation

customer_sales: pyspark.sql.connect.dataframe.DataFrame = [CustomerKey: integer, TotalSpend: double ... 2 more fields]

Table +

	CustomerKey	TotalSpend	TotalOrders	AvgOrderValue
1	19204	539.99	1	539.99
2	21700	1054.4275	1	527.21375
3	29228	49.98	1	49.98
4	13623	156.35	3	26.058333333333334
5	12046	43.6442	2	21.8221
6	15846	4704.3624	2	588.0453
7	28759	103.6042	1	34.534733333333335
8	13289	49.56	2	16.52
9	28836	2443.35	1	2443.35
10	18051	1735.98	1	867.99

Fig: Customer Insights(Lifetime Value) Data Display

```
Aug 22, 2025 (11s)                                                 19
customer_sales.write.mode("overwrite").format("delta").save("abfss://gold@strgaccf0.dfs.core.windows.net/
AdventureWorks_CustomerSales/")

(4) Spark Jobs
```

Fig: Customer Insights(Lifetime Value) Data Save Into Gold Layer

TRANSFORMATION USING DATABRICKS CONTINUED...

Data Aggregation and Storing into Gold Layer Continued...

Fig: Returns Data Load Into DataBricks From Silver

Fig: Returns Data Display

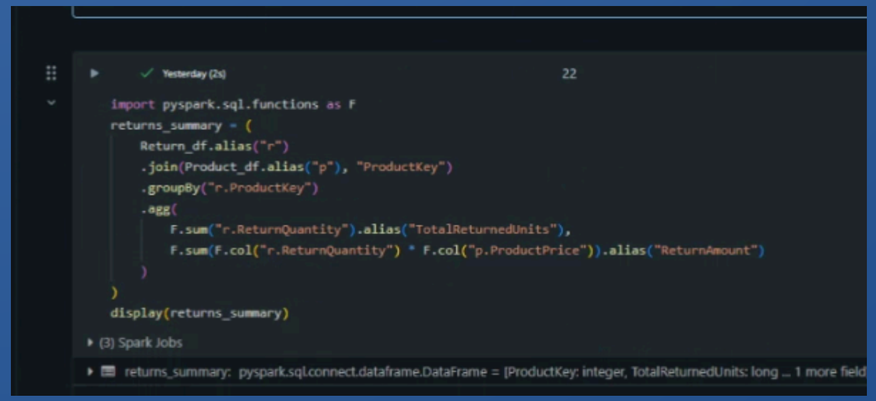


Fig: Returns Rate By Product Aggregation

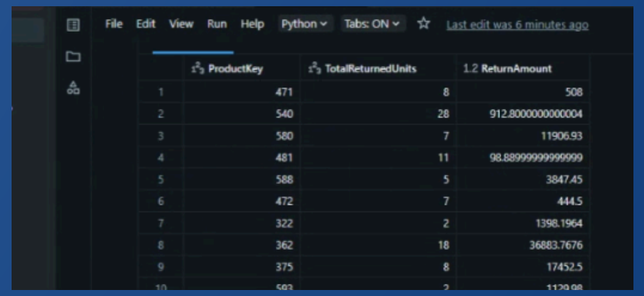


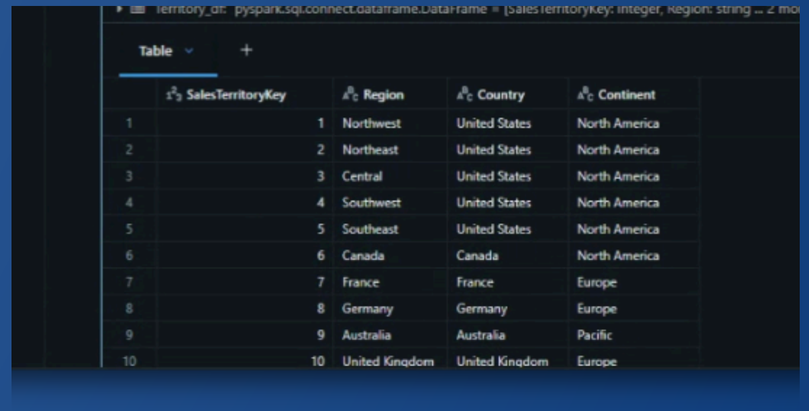
Fig: Returns Rate By Product Data Display

Fig: Returns Rate By Product Data Save Into Gold Layer

TRANSFORMATION USING DATABRICKS CONTINUED...

Data Aggregation and Storing into Gold Layer Continued...

Fig: Territory Data Load Into DataBricks From Silver



Table

	SalesTerritoryKey	Region	Country	Continent
1	1	Northwest	United States	North America
2	2	Northeast	United States	North America
3	3	Central	United States	North America
4	4	Southwest	United States	North America
5	5	Southeast	United States	North America
6	6	Canada	Canada	North America
7	7	France	France	Europe
8	8	Germany	Germany	Europe
9	9	Australia	Australia	Pacific
10	10	United Kingdom	United Kingdom	Europe

Fig: Territory Data Display

TRANSFORMATION USING DATABRICKS CONTINUED...

Data Aggregation and Storing into Gold Layer Continued...

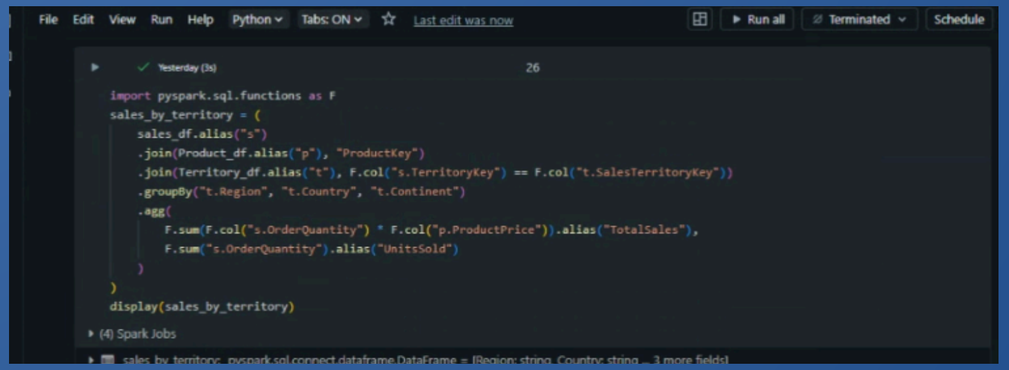


Fig: Sales By Territory Aggregation

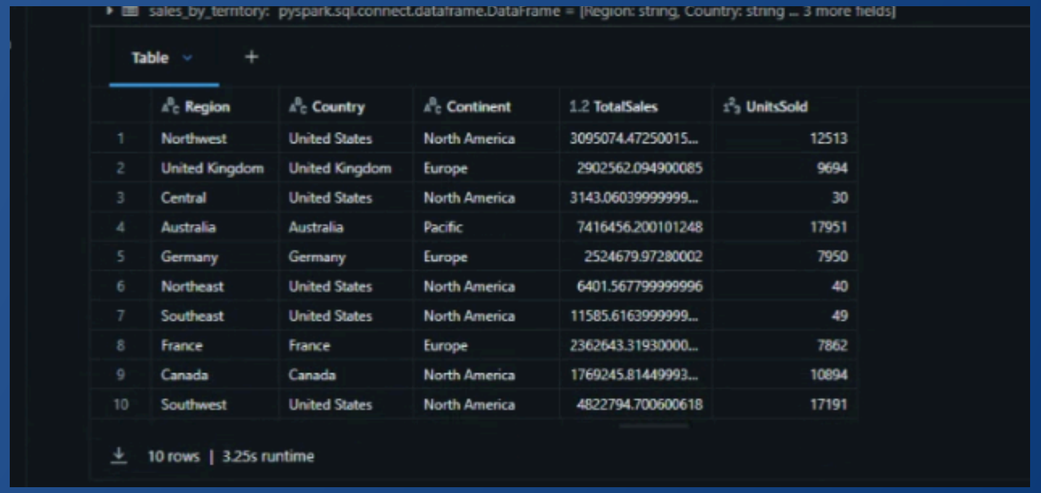


Fig: Sales By Territory Data Display

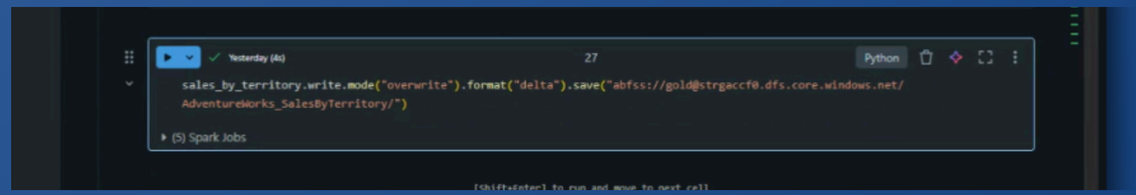
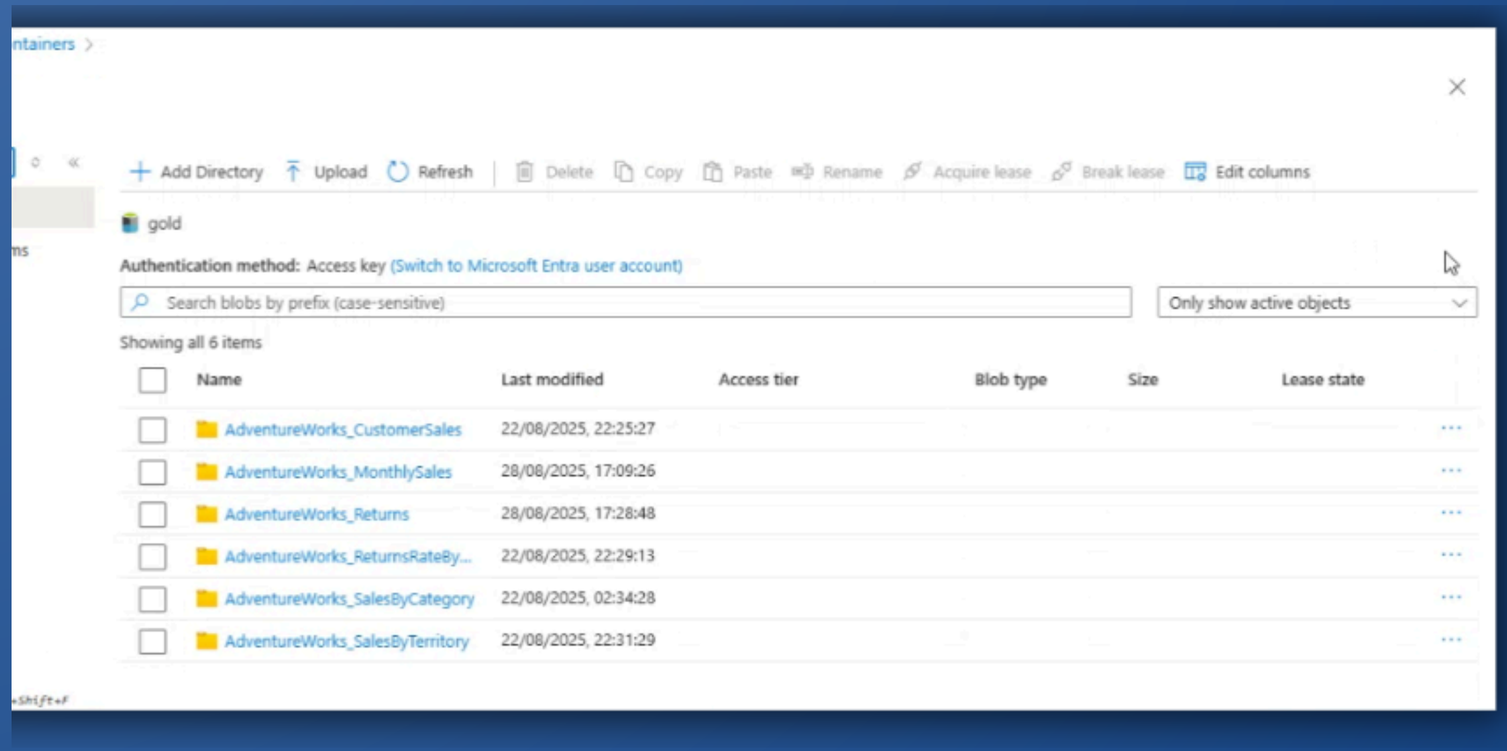


Fig: Sales By Territory Data Save Into Gold Layer

TRANSFORMATION USING DATABRICKS CONTINUED...

Successful Data Store : Silver to Gold



Successful Data Store Into Gold Layer

FUTURE SCOPE

As a future scope of this project, I plan to connect the Gold layer with Power BI to generate meaningful business insights and interactive reports. This will help in visualizing transformed data and supporting data-driven decision-making.

CONCLUSION

This project successfully demonstrates the power of Azure Data Engineering technologies in building a scalable, secure, and automated data pipeline. By integrating Azure Data Factory for ingestion and orchestration, Databricks with PySpark for transformation, and Delta Lake for optimized storage, the solution ensures efficient handling of raw data from diverse sources. The architecture not only improves data quality and performance, but also enables seamless analytics and business intelligence, making it a robust foundation for real-world data-driven decision-making.

THANK YOU