

DENIS PROJECT

Data Engineering using Medallion Architecture

By Shanti Jogi

Introduction:

The DENIS Project is a data engineering project that follows the Medallion Architecture to transform and curate data for reporting. The project was implemented using Azure Data Factory (ADF), Azure Data Lake Storage (ADLS), Azure SQL Database, and Power BI. The primary goal of the project is to ingest, transform, and analyse data from various sources, including Blob Storage, ADLS, MySQL, and Cosmos DB, and ultimately provide a final analytical dataset in the Gold layer.

Project Workflow Overview

The project follows a 3-layer architecture based on the Medallion architecture:

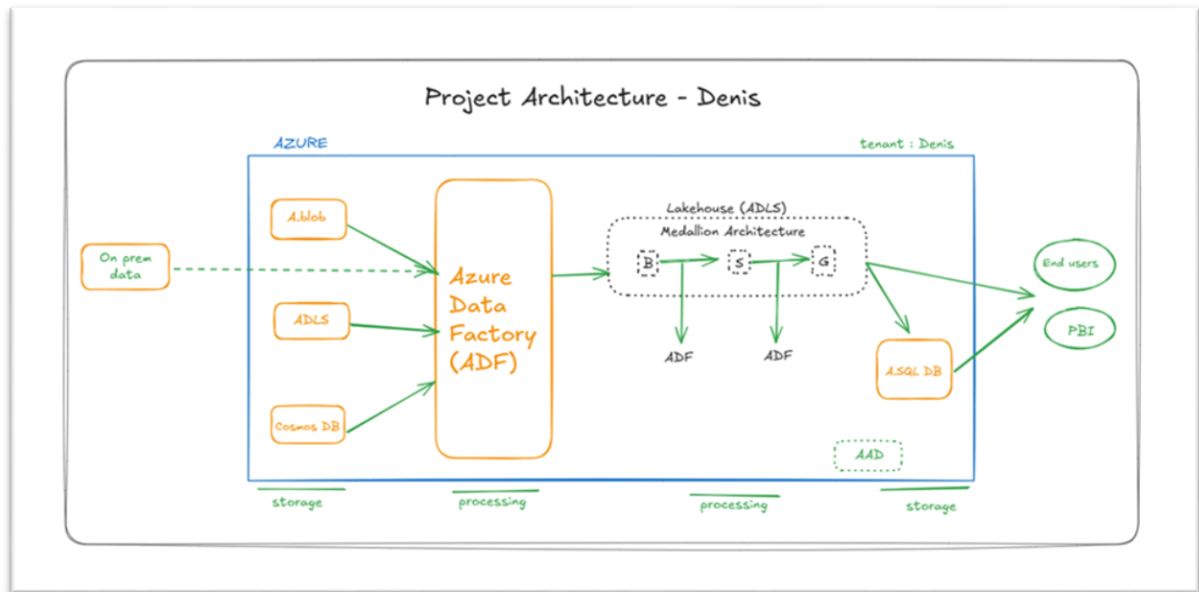
1. **Bronze Layer:** Raw, untransformed data.
2. **Silver Layer:** Cleaned and structured data with business logic applied.
3. **Gold Layer:** Curated, report-ready dataset that combines all data into a single analytical table.

Each layer is implemented using Azure services, particularly Azure Data Factory, Azure Data Lake Storage, and Azure SQL Database.

Architecture and Design

The data pipeline architecture followed the **Medallion Architecture**, with three key stages:

1. **Bronze Layer:** This layer ingested raw data from various sources such as **Blob Storage**, **Azure Data Lake Storage (ADLS)**, **Cosmos DB**, and **On-Prem MySQL**. The data was stored in its raw format for future processing.
2. **Silver Layer:** This layer transformed the data by cleaning, structuring, and enriching it. Transformations included trimming, creating new columns, adjusting data types, and selecting only necessary columns.
3. **Gold Layer:** The final analytical layer that joined fact and dimension data to create a single, optimized dataset for reporting and business intelligence.

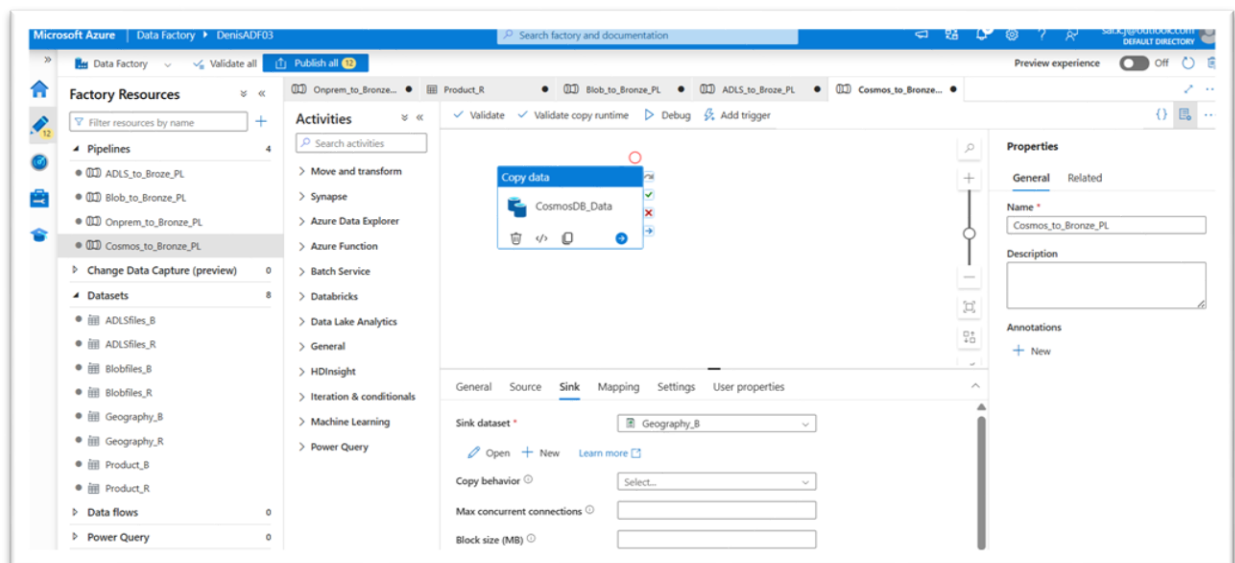


Bronze Layer

The **Bronze Layer** is where raw data was ingested from multiple sources into Azure Data Lake Storage. This raw data is stored as-is without any transformations.

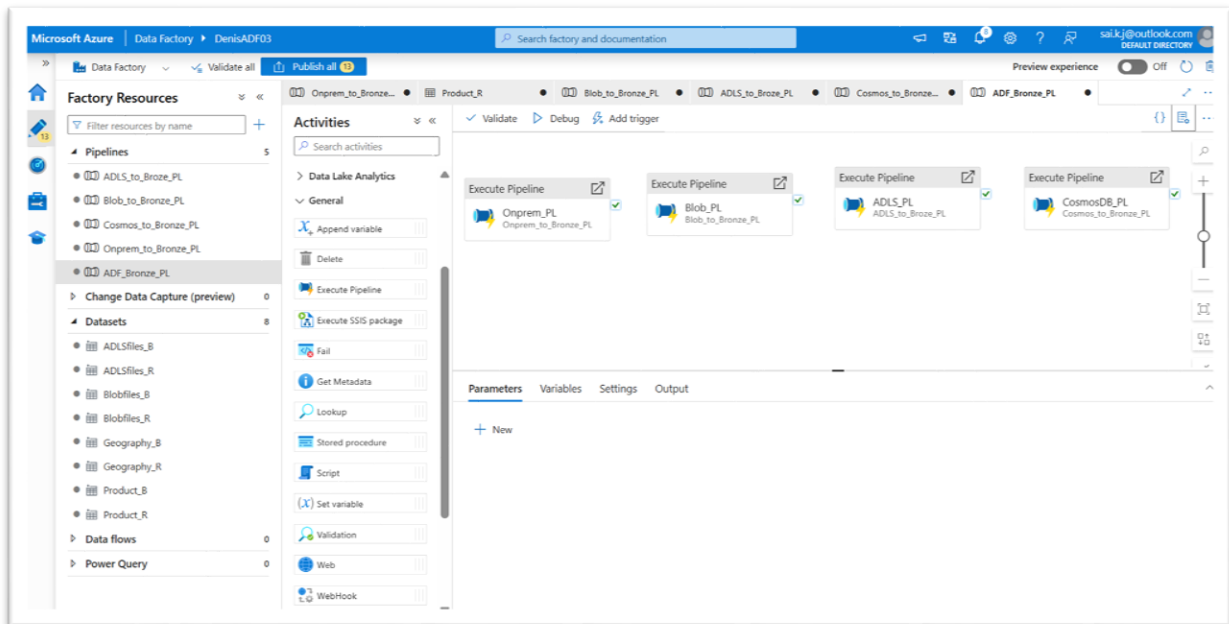
Sources Ingested:

- **Blob Storage:** Data files in **CSV** and **Excel** formats.
- **Azure Data Lake Storage (ADLS):** Additional CSV and Excel files.
- **Cosmos DB:** Data in **JSON** format.
- **On-Prem MySQL:** Data from an on-premises database.



Process:

- Data was ingested into separate folders for each source: /bronze/blob/, /bronze/adls/, /bronze/cosmos/, and /bronze/mysql/.
- Separate **Azure Data Factory (ADF)** pipelines were created for each source, ensuring that the data was copied into the Bronze layer efficiently.



Pipeline triggered

Activity runs

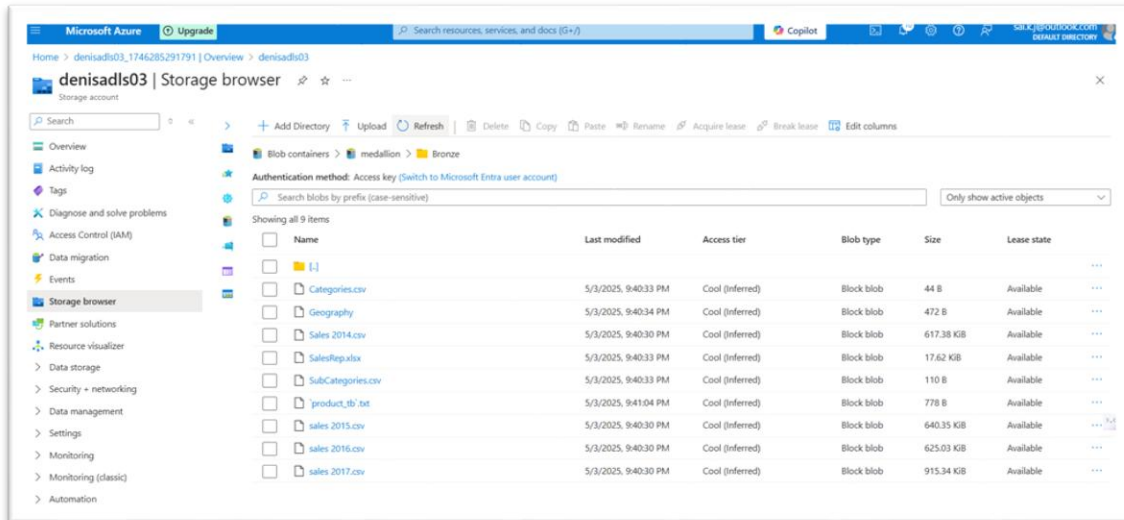
Pipeline run ID: 88ae8937-2e85-4593-8972-5f3487d62bb7

All status: Monitor in Azure Metrics Export to CSV

Showing 1 - 4 items

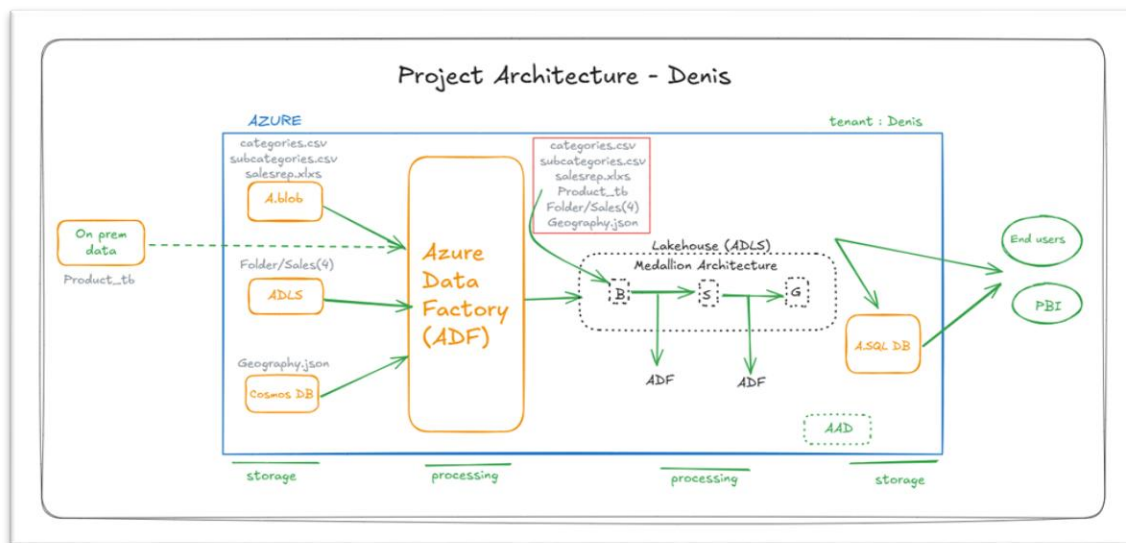
Activity name	Activity st...	Activit...	Run start	Duration	Integration runtime	User prop...	Activity run ID
CosmosDB_PL	Succeeded	Execute Pipelin	5/3/2025, 9:28:04 PM	25s			b90710db-21c0-4160-9e62-e5e
Onprem_PL	Succeeded	Execute Pipelin	5/3/2025, 9:28:04 PM	44s			525e983e-3344-4e19-8729-d23
ADLS_PL	Succeeded	Execute Pipelin	5/3/2025, 9:28:04 PM	17s			3b2f5b31-ccb4-40c8-90eb-8a3i
Blob_PL	Succeeded	Execute Pipelin	5/3/2025, 9:28:04 PM	21s			964840d6-e404-4977-9e29-96f

Data Moved to bronze Layer:



Name	Last modified	Access tier	Blob type	Size	Lease state
Categories.csv	5/3/2025, 9:40:33 PM	Cool (Inferred)	Block blob	44 B	Available
Geography	5/3/2025, 9:40:34 PM	Cool (Inferred)	Block blob	472 B	Available
Sales 2014.csv	5/3/2025, 9:40:30 PM	Cool (Inferred)	Block blob	617.38 KB	Available
SalesRep.xlsx	5/3/2025, 9:40:33 PM	Cool (Inferred)	Block blob	17.62 KB	Available
SubCategories.csv	5/3/2025, 9:40:33 PM	Cool (Inferred)	Block blob	110 B	Available
'product.tb'.txt	5/3/2025, 9:41:04 PM	Cool (Inferred)	Block blob	778 B	Available
sales 2015.csv	5/3/2025, 9:40:30 PM	Cool (Inferred)	Block blob	640.35 KB	Available
sales 2016.csv	5/3/2025, 9:40:30 PM	Cool (Inferred)	Block blob	625.03 KB	Available
sales 2017.csv	5/3/2025, 9:40:30 PM	Cool (Inferred)	Block blob	915.34 KB	Available

Bronze Layer Successfully Created



Silver Layer

The Silver Layer involved transforming the data from the Bronze layer into a cleaner, structured form, making it ready for analytical processing.

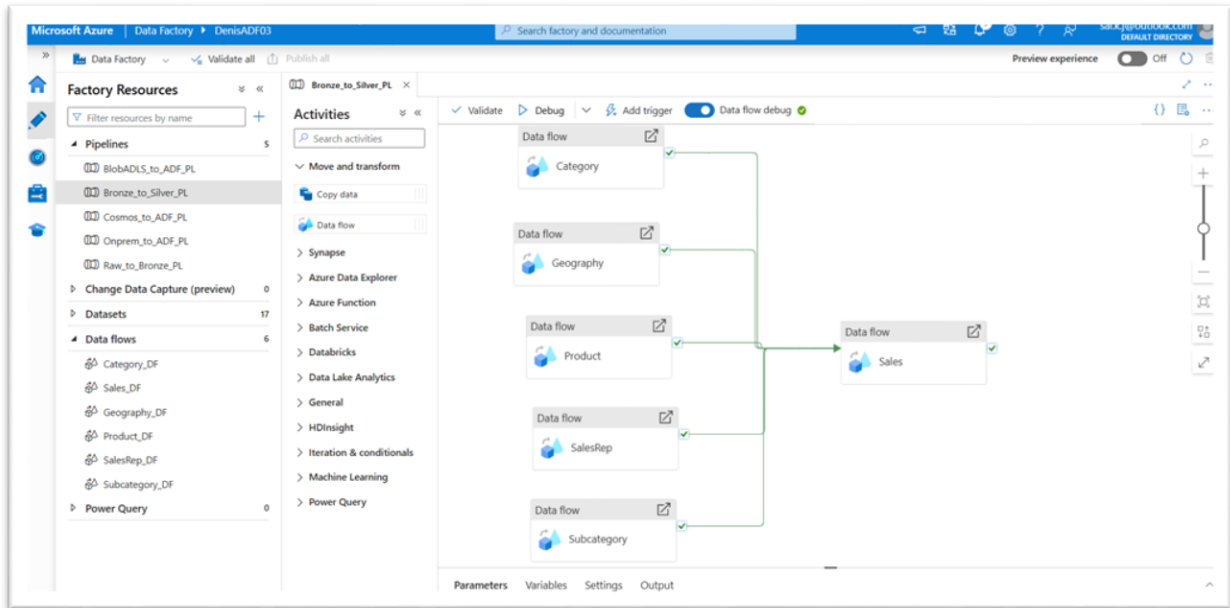
Dataflows:

- 6 Dataflows were created in Azure Data Factory:
 - Dimension Data: Transformation of customer, product, and other dimension data.
 - Fact Data: Sales data transformation.

Shanti Jogi DE

3. Additional dataflows cleaned and transformed necessary data for reporting.

- Transformations Applied:
 - Data Cleaning: Removed unnecessary columns.
 - Derived Columns: Created new columns based on existing data.
 - Data Type Adjustments: Set appropriate data types for each column.
 - Regex Operations: Cleaned and trimmed values.

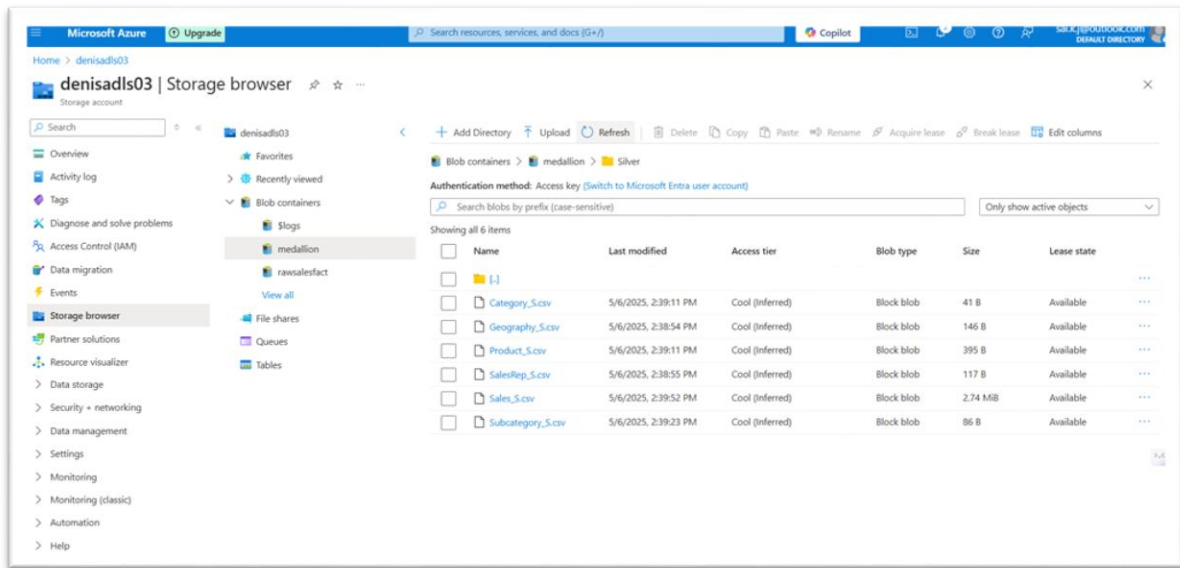


Pipeline triggered

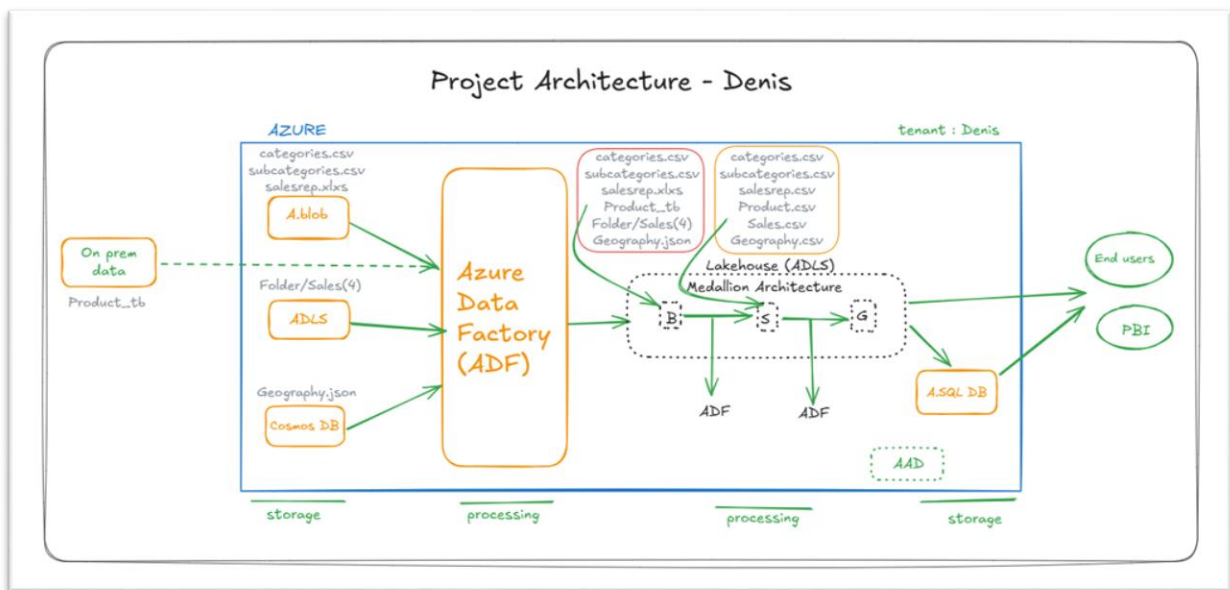
The screenshot displays the Microsoft Azure Data Factory console for the 'Bronze_to_Silver_PL' pipeline, specifically the 'Activity runs' section. The interface shows a list of activity runs for the pipeline, including details such as activity name, status, run start time, duration, integration runtime, and user properties. The table shows five activity runs, all of which are 'Succeeded'.

Activity name	Activity st...	Activit...	Run start	Duration	Integration runtime	User prop...	Activity run ID
Subcategory	Succeeded	Data flow	5/4/2025, 5:16:11 PM	3m 23s	AutoResolveIntegrationRuntime (East US)		22a49090-3298-4623-b4b7-69f
SalesRep	Succeeded	Data flow	5/4/2025, 5:16:11 PM	3m 22s	AutoResolveIntegrationRuntime (East US)		4becbd90-fa4a-4075-b689-a14
Geography	Succeeded	Data flow	5/4/2025, 5:16:11 PM	3m 18s	AutoResolveIntegrationRuntime (East US)		fa46bc9e-e738-4ac1-af85-569e
Category	Succeeded	Data flow	5/4/2025, 5:16:11 PM	3m 24s	AutoResolveIntegrationRuntime (East US)		572d0c28-d5d1-486d-8f4e-bb0
Product	Succeeded	Data flow	5/4/2025, 5:16:11 PM	3m 13s	AutoResolveIntegrationRuntime (East US)		9227b7df-e025-4cab-9ee7-1ccf

Successfully data moved to Silver Layer in one formatted files i.e. csv



Silver Layer successfully created



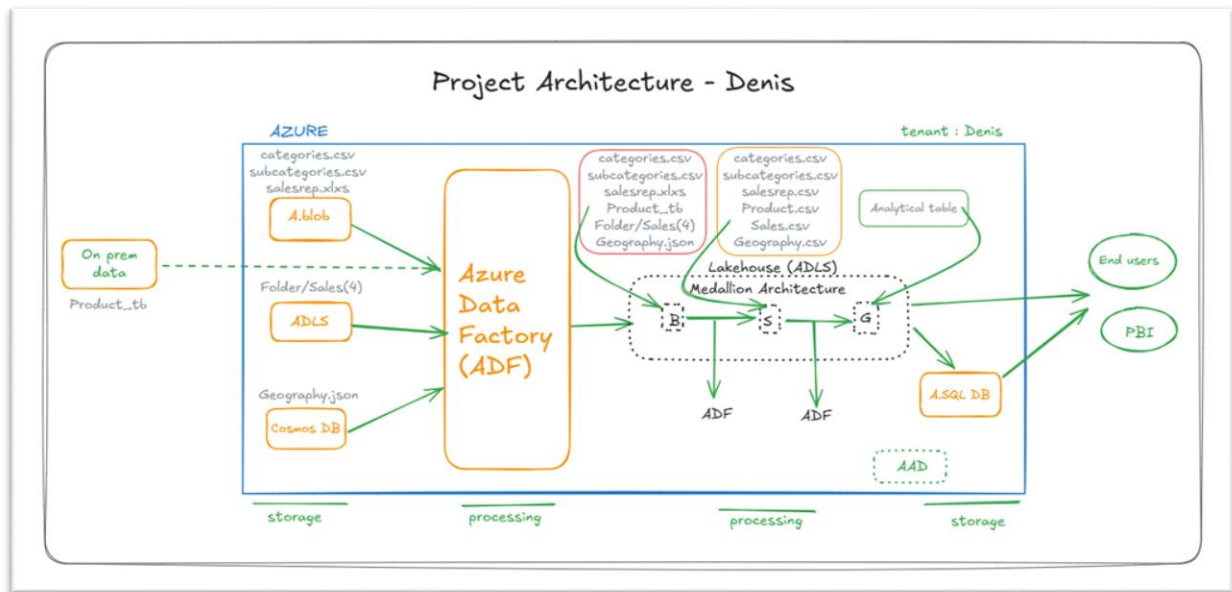
Gold Layer

The Gold Layer contained the final dataset prepared for reporting and business analysis.

Final Dataflow:

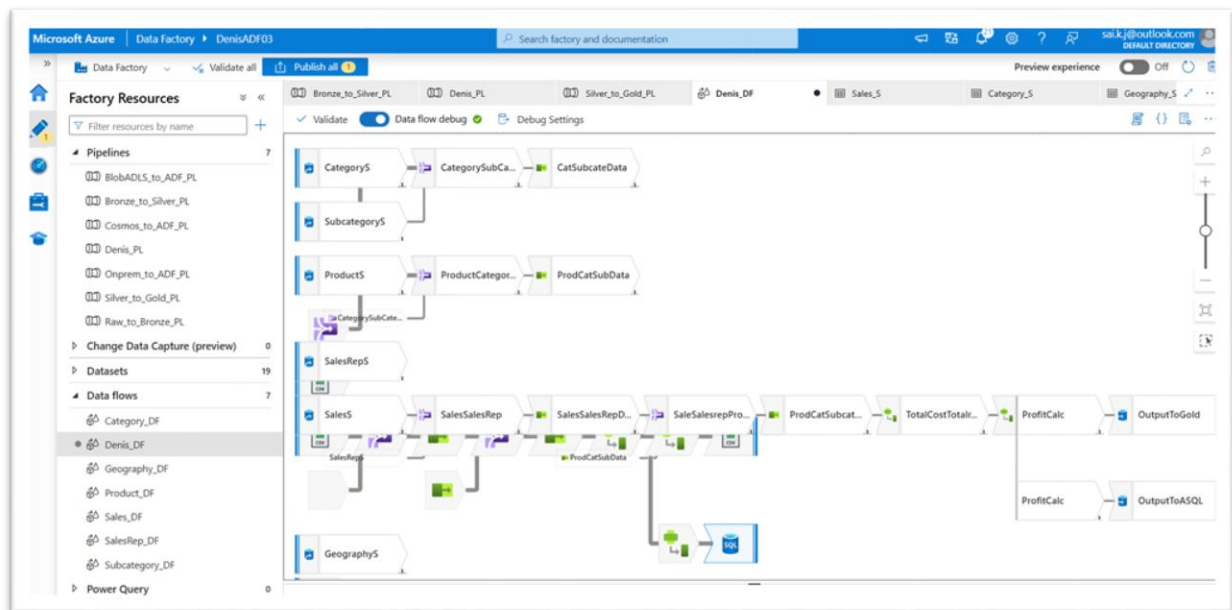
- A single dataflow was created to combine all dimension and fact data.
- Joins were made between the transformed sales data (Silver) and the dimension data to create a complete analytical dataset.

- Unnecessary columns were removed to ensure the dataset was optimized for reporting.



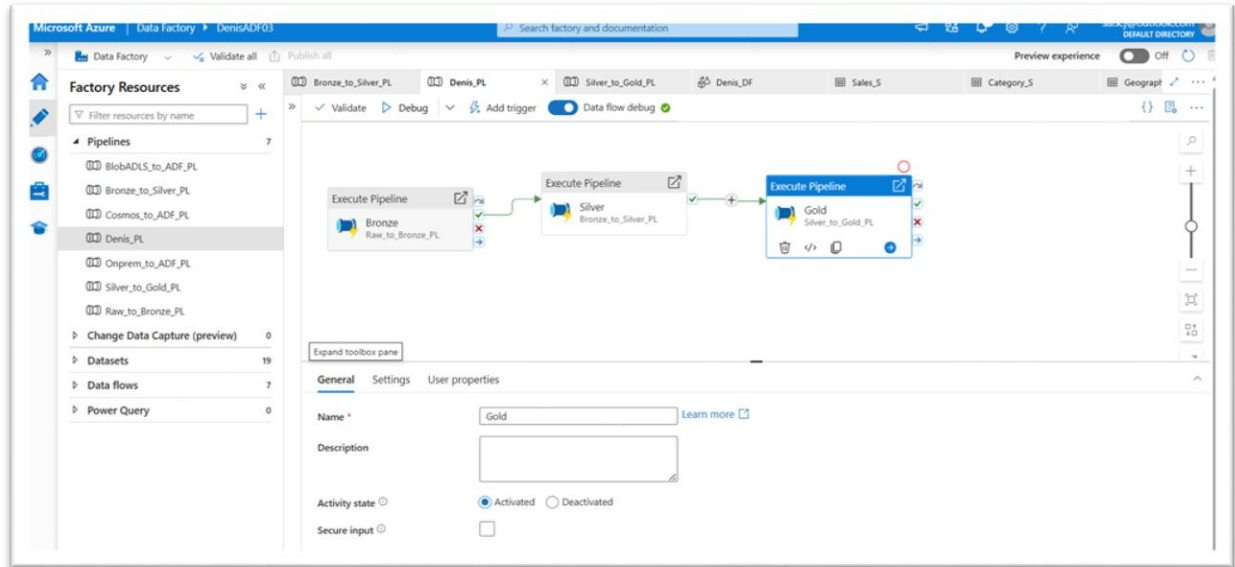
Execution:

- The final dataset was stored in Azure Data Lake Storage (Gold) and Azure SQL Database for Power BI integration.

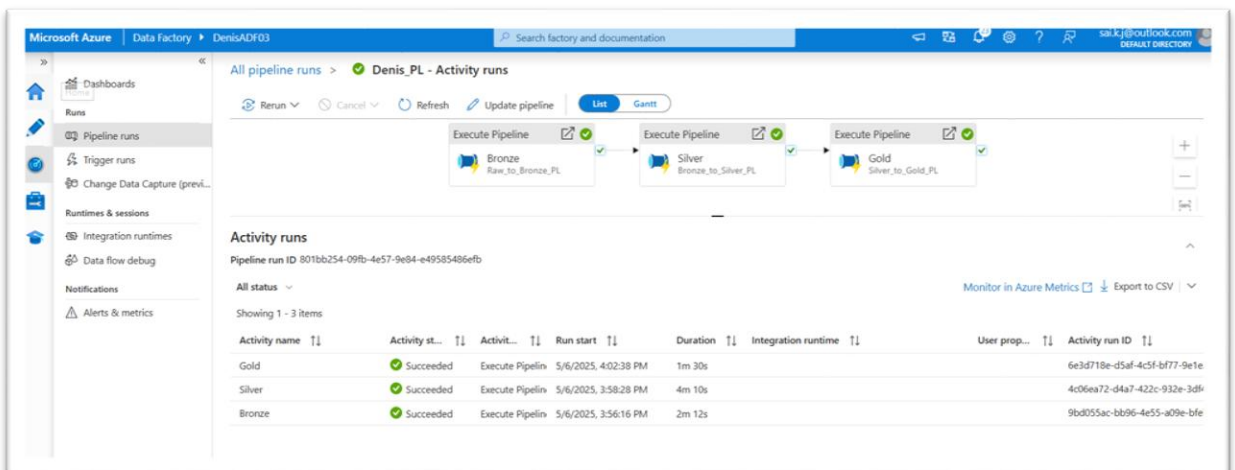


Here I already connected this dataflow to silver to gold layer pipeline and executed this pipeline in Master pipeline i.e. Denis_PL

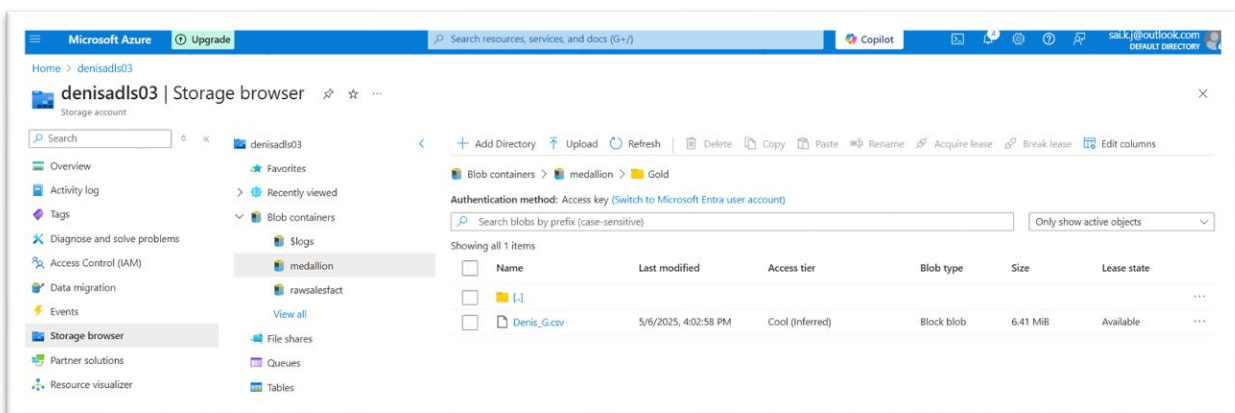
Shanti Jogi DE



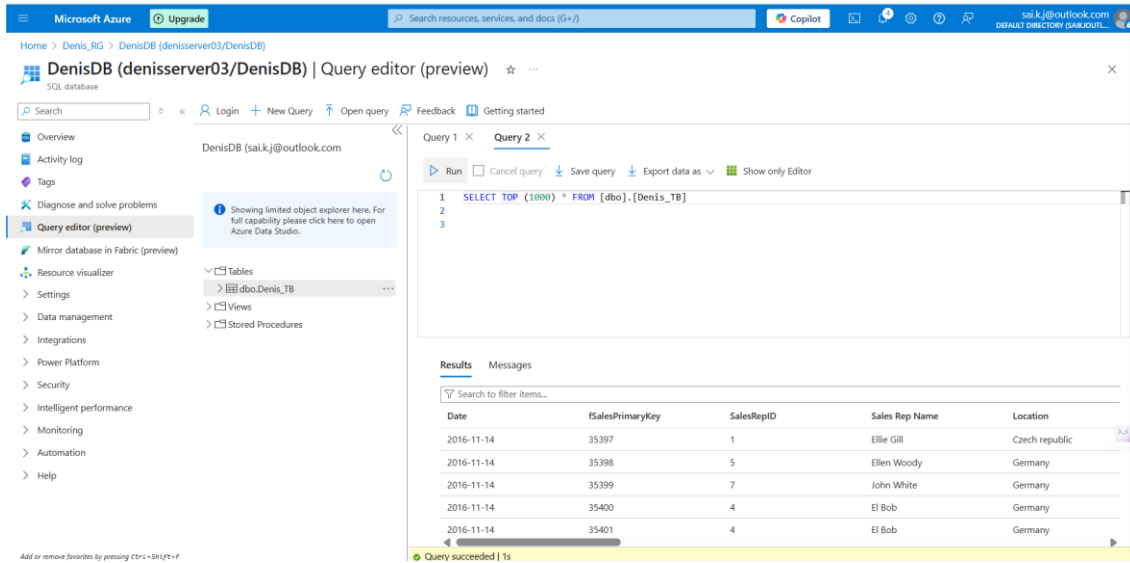
Triggered Pipeline



Data Moved to Gold Layer



Also data Moved to Azure SQL Database

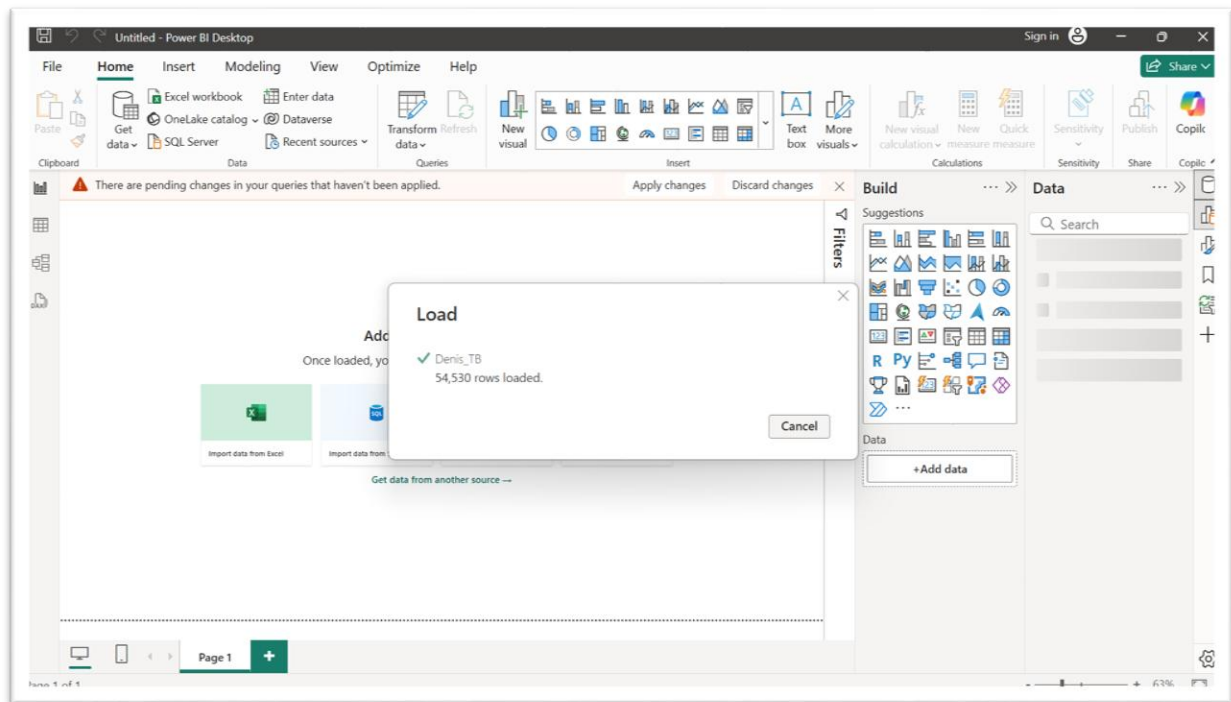


The screenshot shows the Microsoft Azure portal interface for the 'DenisDB (denisserver03/DenisDB)' SQL database. The 'Query editor (preview)' is open, displaying a query: `SELECT TOP (1000) * FROM [dbo].[Denis_TB]`. The results are shown in a table with the following data:

Date	fSalesPrimarykey	SalesRepID	Sales Rep Name	Location
2016-11-14	35397	1	Ellie Gill	Czech republic
2016-11-14	35398	5	Ellen Woody	Germany
2016-11-14	35399	7	John White	Germany
2016-11-14	35400	4	El Bob	Germany
2016-11-14	35401	4	El Bob	Germany

The status bar at the bottom indicates 'Query succeeded | 1s'.

Connected this data to Power Bi via Azure SQL Database Server



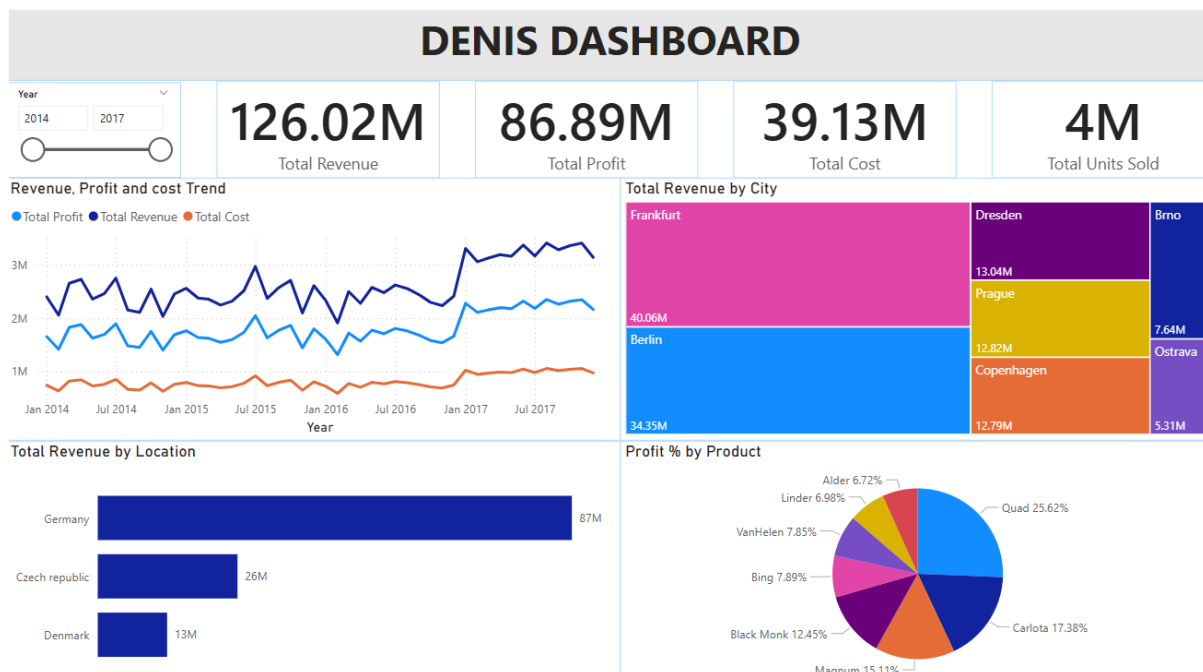
The screenshot shows the Power BI Desktop interface. A 'Load' dialog box is open, indicating that 54,530 rows have been loaded from the 'Denis_TB' table. The dialog box has a 'Cancel' button. The background shows the Power BI Desktop ribbon with the 'Home' tab selected, and the 'Data' pane on the right side.

Shanti Jogi DE

Shanti Jogi DE

Successful connected and loaded the data

Created Denis Dashboard:



Conclusion

The DENIS Project successfully built a scalable, automated data pipeline using Azure Data Factory and Power BI, following the Medallion Architecture. The project met the client's requirements for data ingestion, transformation, and reporting. The final Gold layer dataset provides clean, structured data for business insights and decision-making.