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:

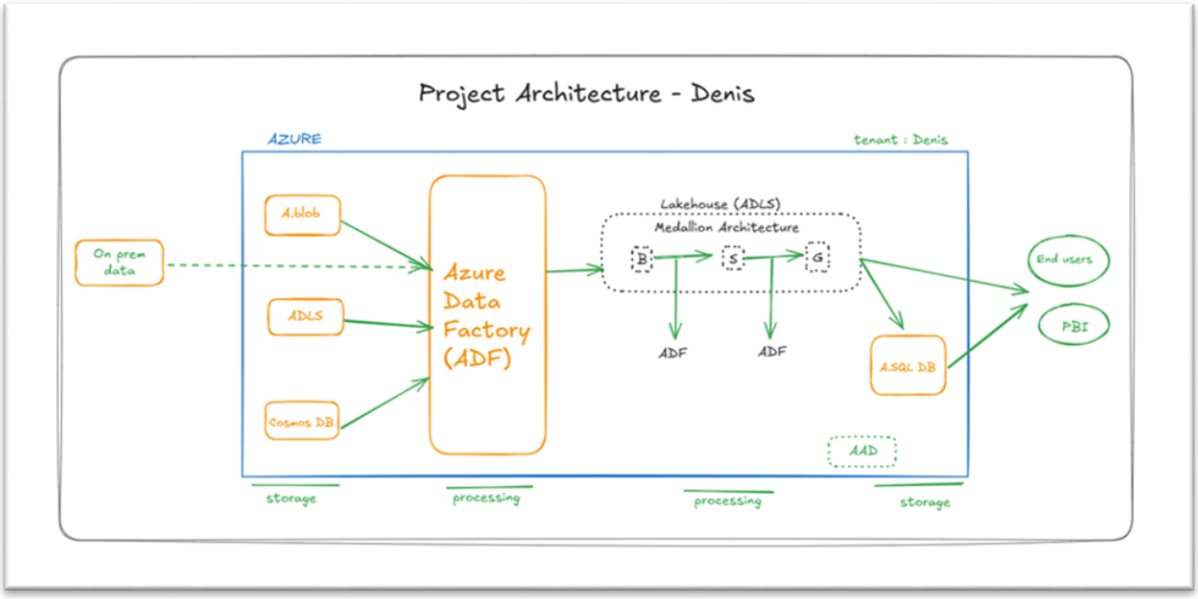
1. ***Bronze Layer***: Raw, untransformed data.
2. ***Silver Layer***: Cleaned and structured data with business logic applied.
3. ***Gold Layer***: Curated, report-ready dataset that combines all data into a single analytical table.

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.

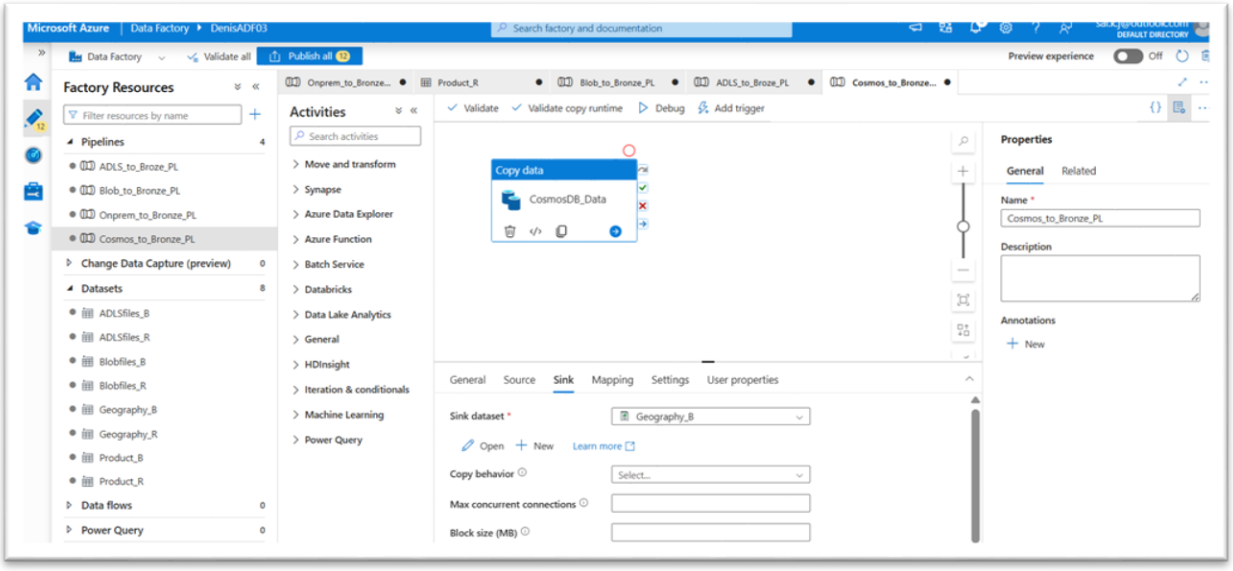


Bronze Layer

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.

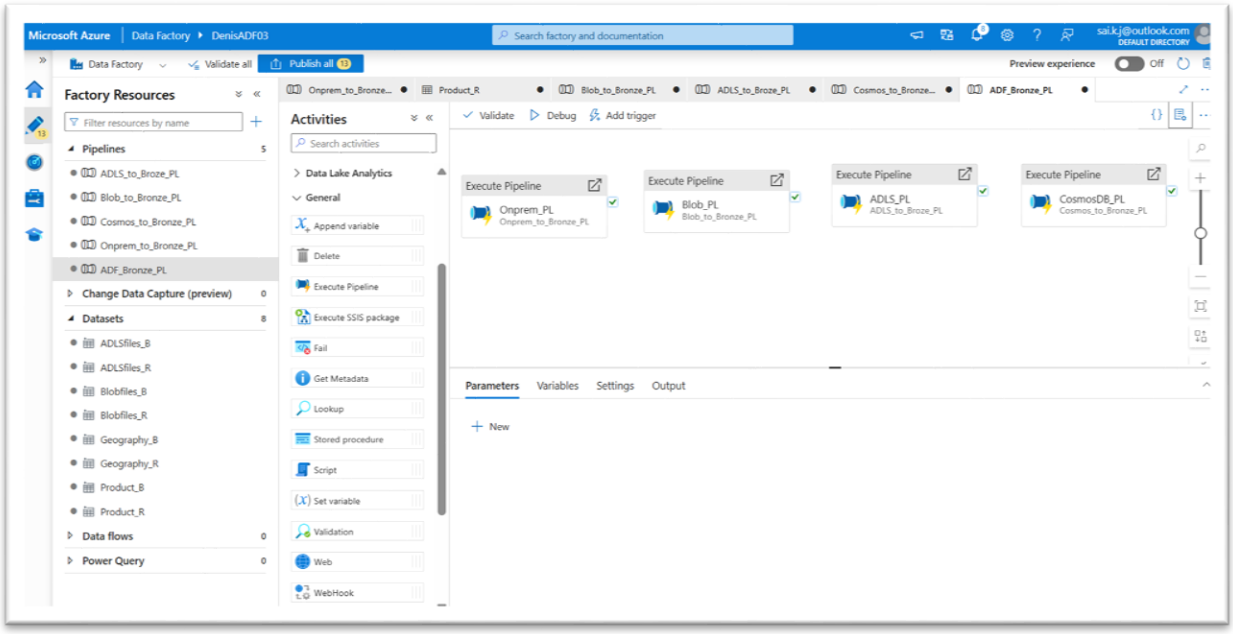
Sources Ingested:

- **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.

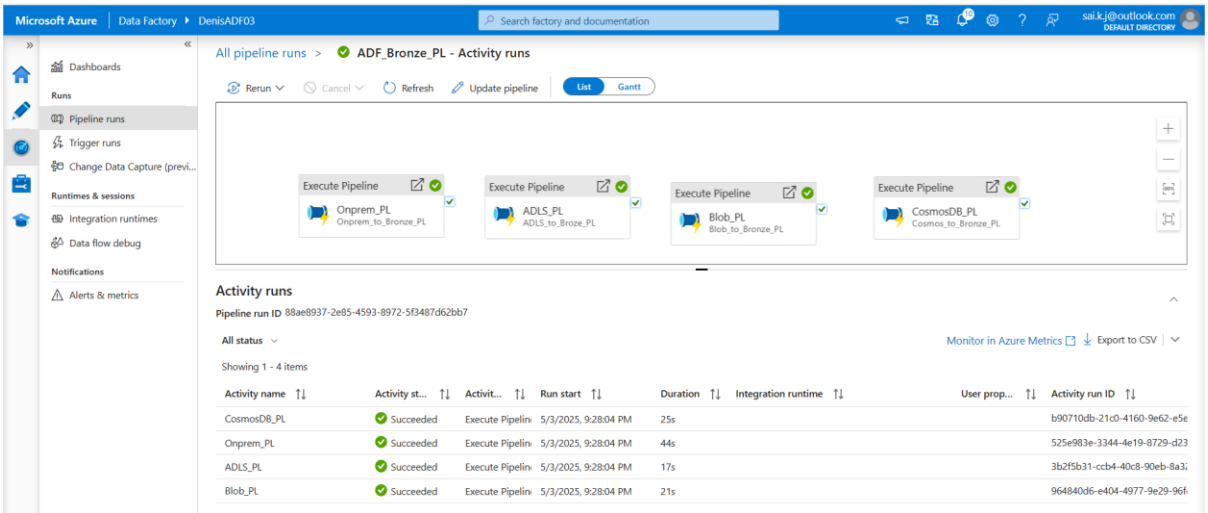


Process:

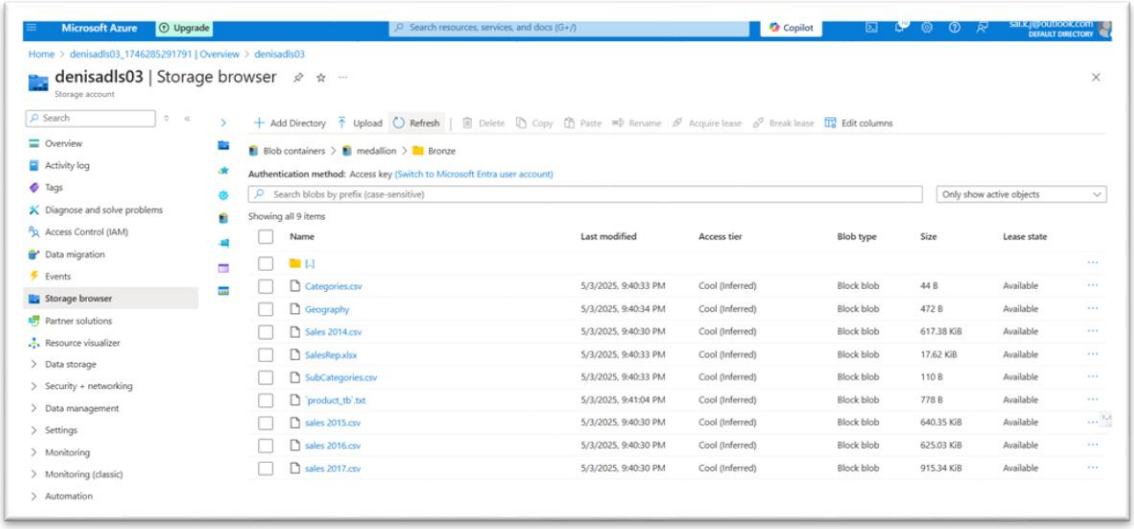
- Data was ingested into separate folders for each source: /bronze/blob/, /bronze/adls/, /bronze/cosmos/, and /bronze/mysql/.
- Separate **Azure Data Factory (ADF)** pipelines were created for each source, ensuring that the data was copied into the Bronze layer efficiently.



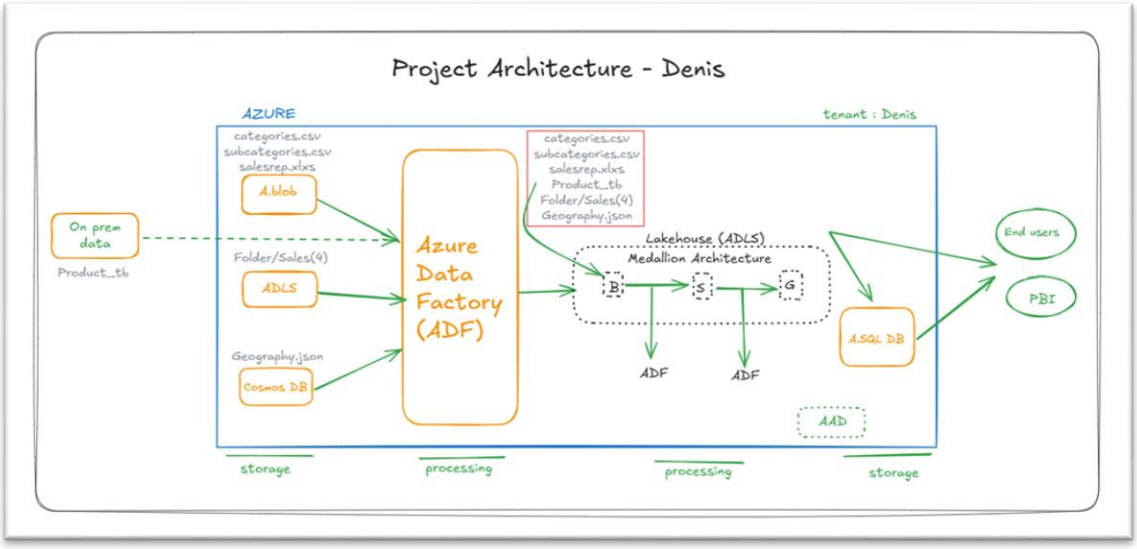
Pipeline triggered



Data Moved to bronze Layer:



Bronze Layer Successfully Created



Silver Layer

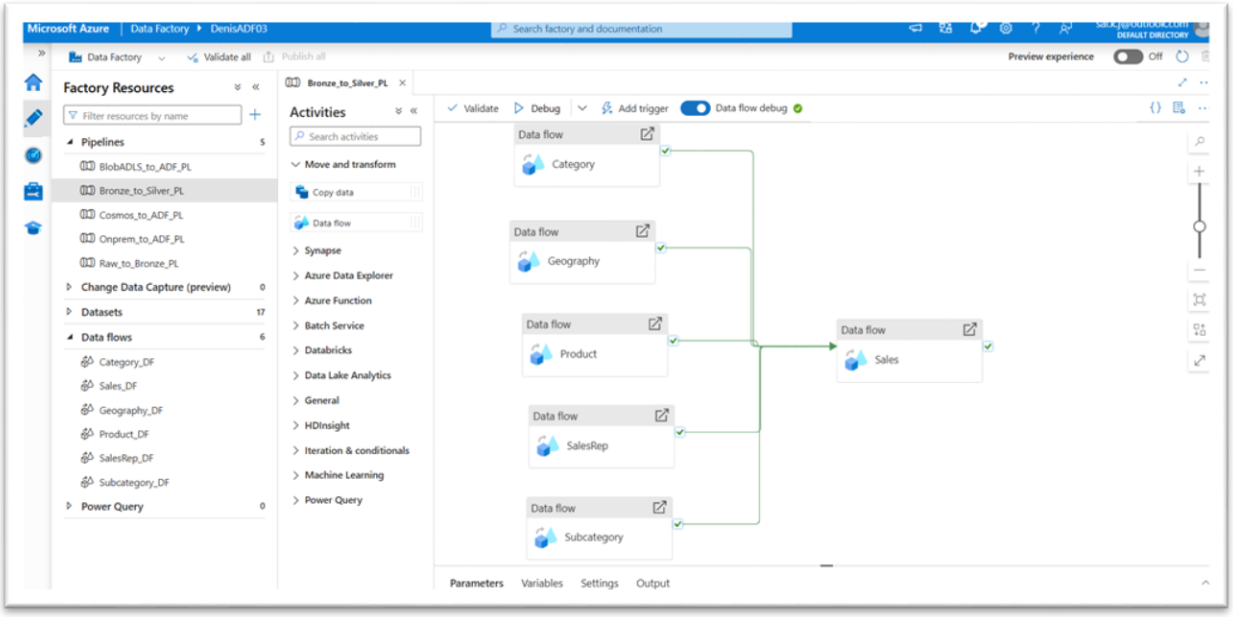
The Silver Layer involved transforming the data from the Bronze layer into a cleaner, structured form, making it ready for analytical processing.

Dataflows:

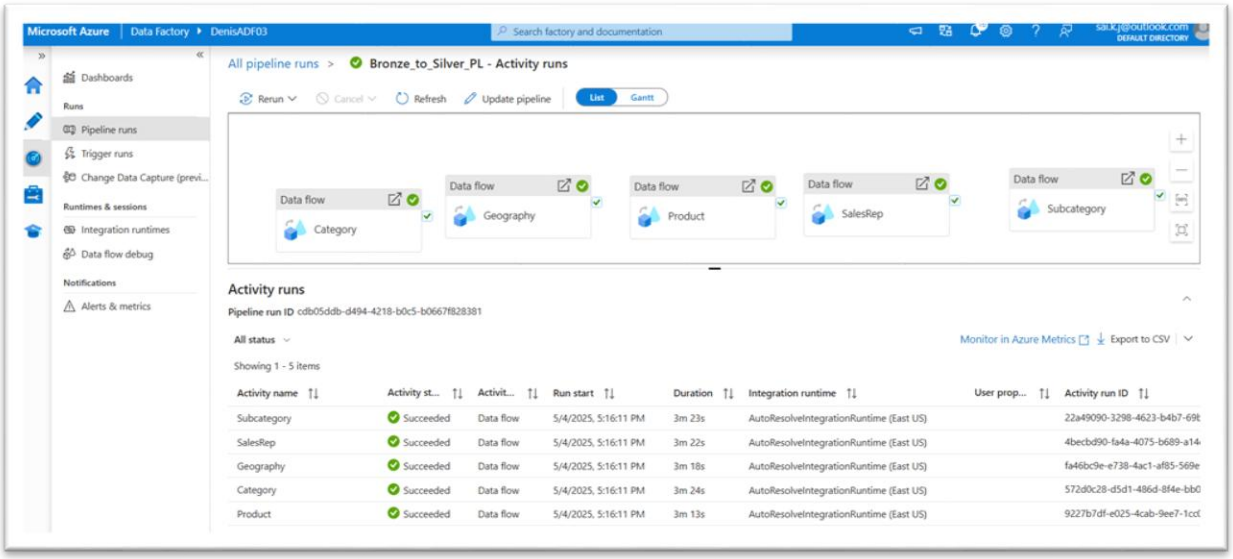
- 6 Dataflows were created in Azure Data Factory:
    1. Dimension Data: Transformation of customer, product, and other dimension data.
    2. Fact Data: Sales data transformation.

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.

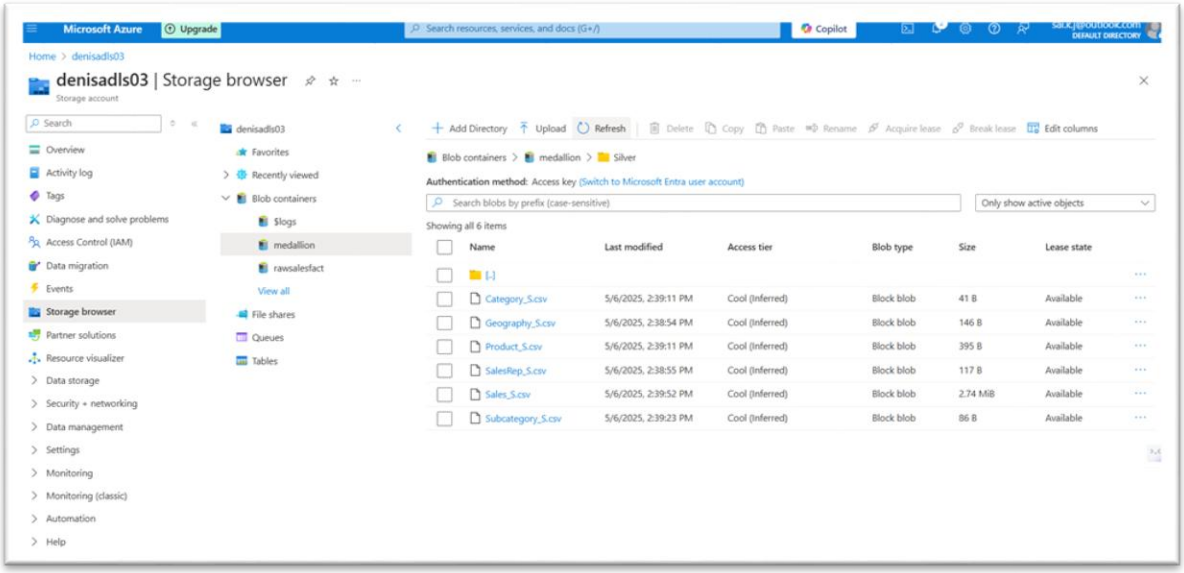


Pipeline triggered

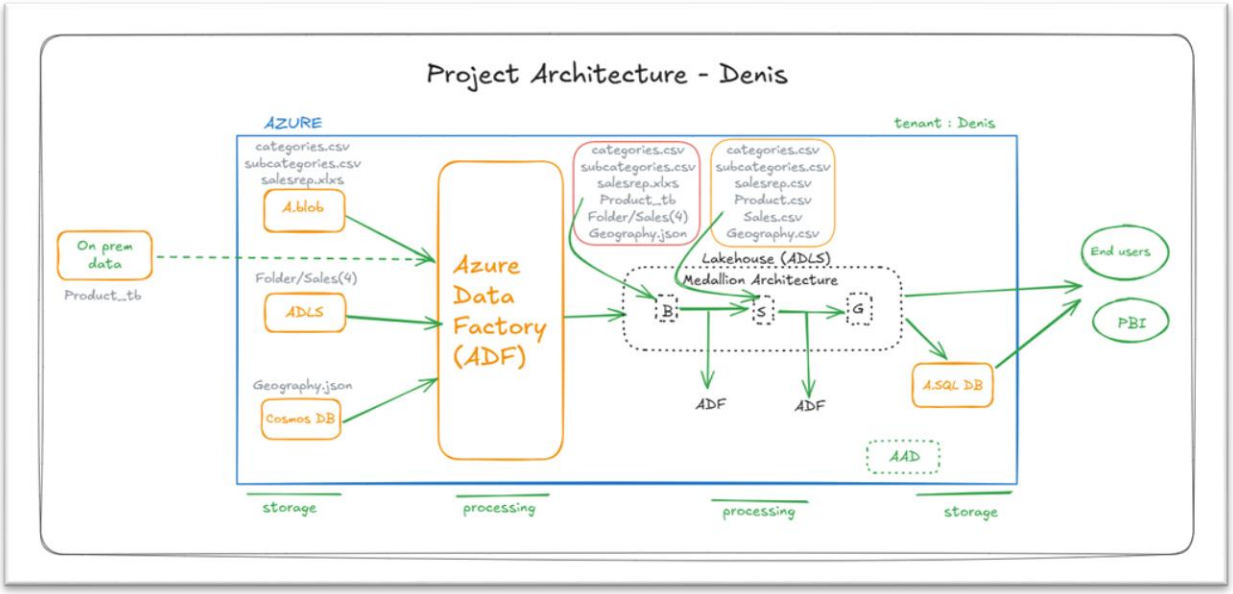


Shanti Jogi DE

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



Silver Layer successfully created



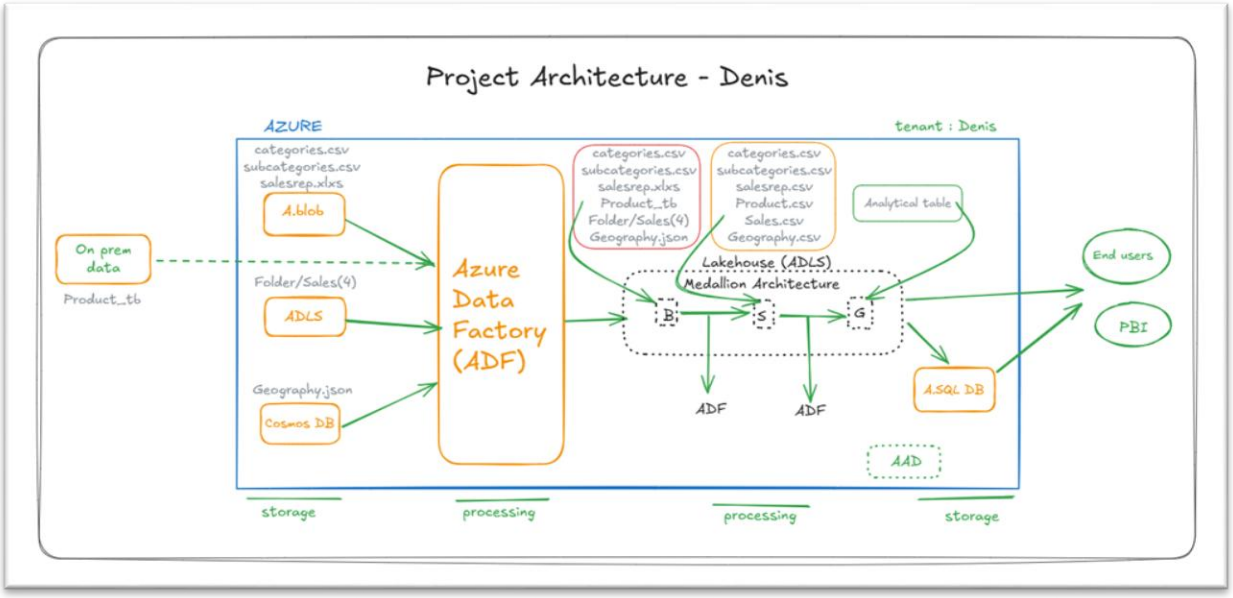
Gold Layer

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

Final Dataflow:

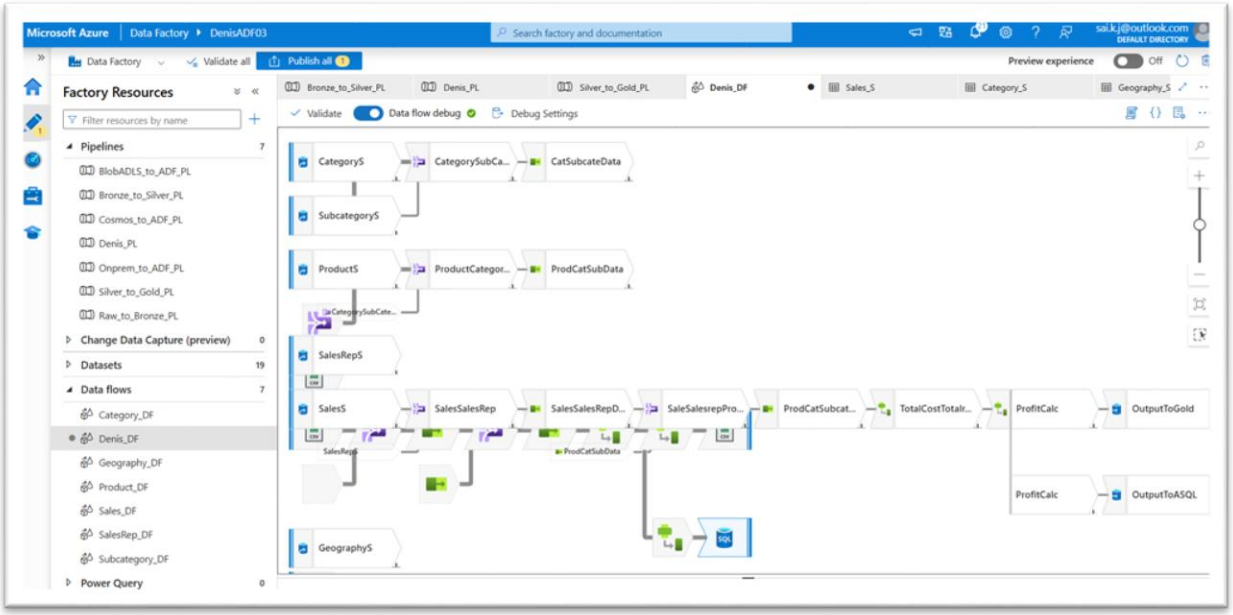
- A single dataflow was created to combine all dimension and fact data.
- Joins were made between the transformed sales data (Silver) and the dimension data to create a complete analytical dataset.

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



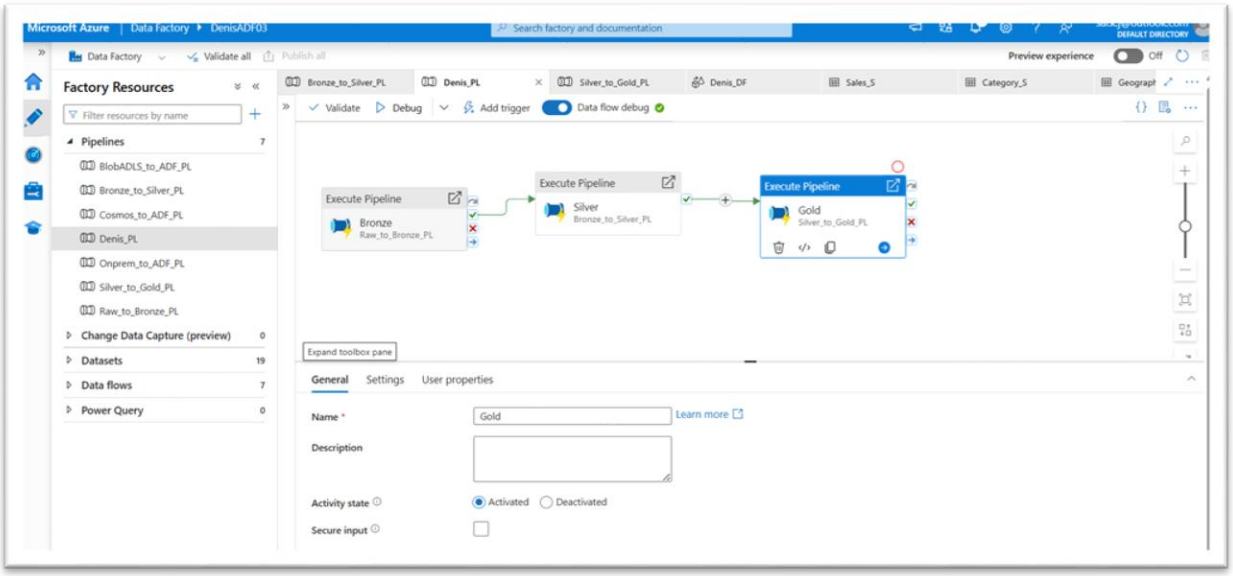
Execution:

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

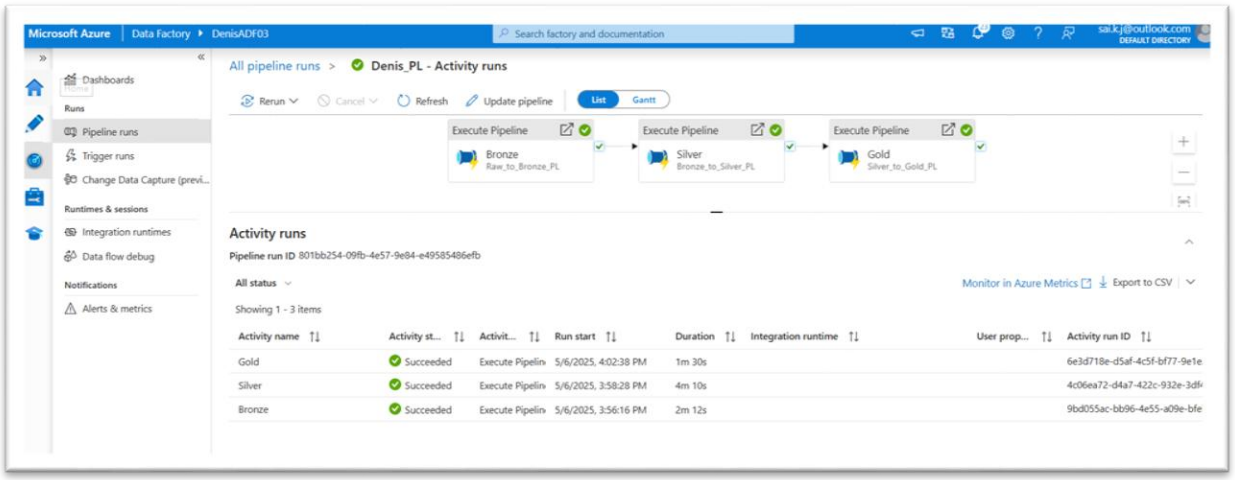


Here I already connected this dataflow to silver to gold layer pipeline and executed this pipeline in Master pipeline i.e. Denis_PL

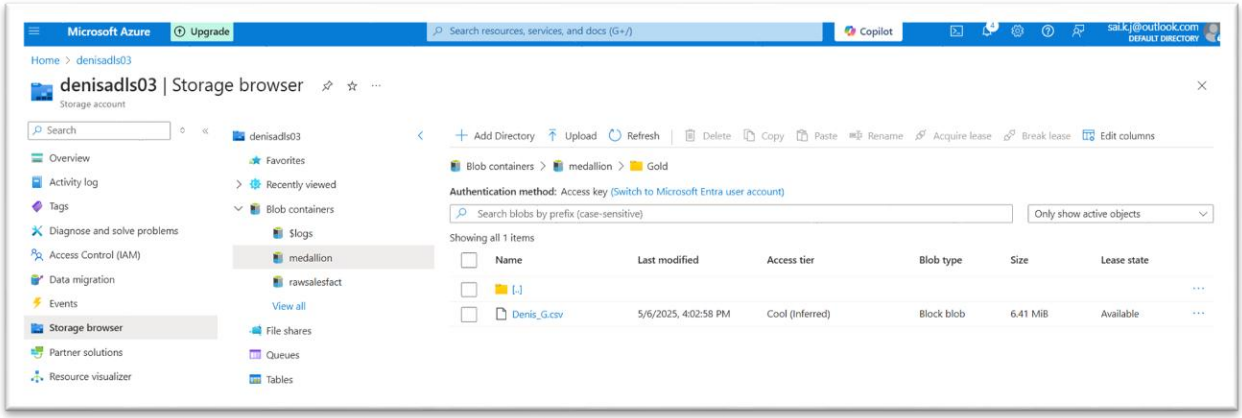
Shanti Jogi DE



Triggered Pipeline

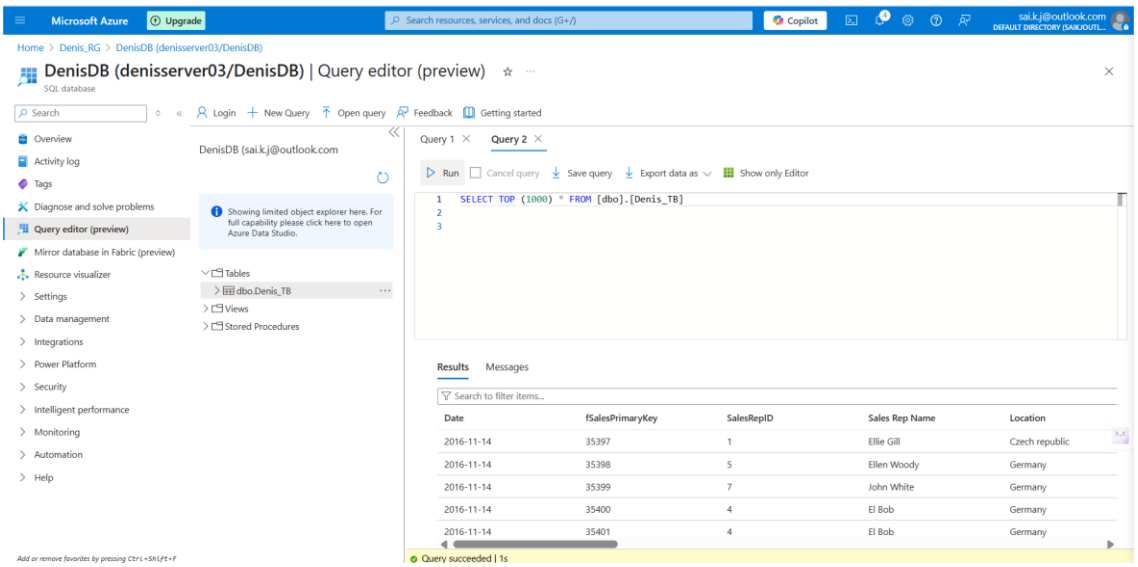


Data Moved to Gold Layer

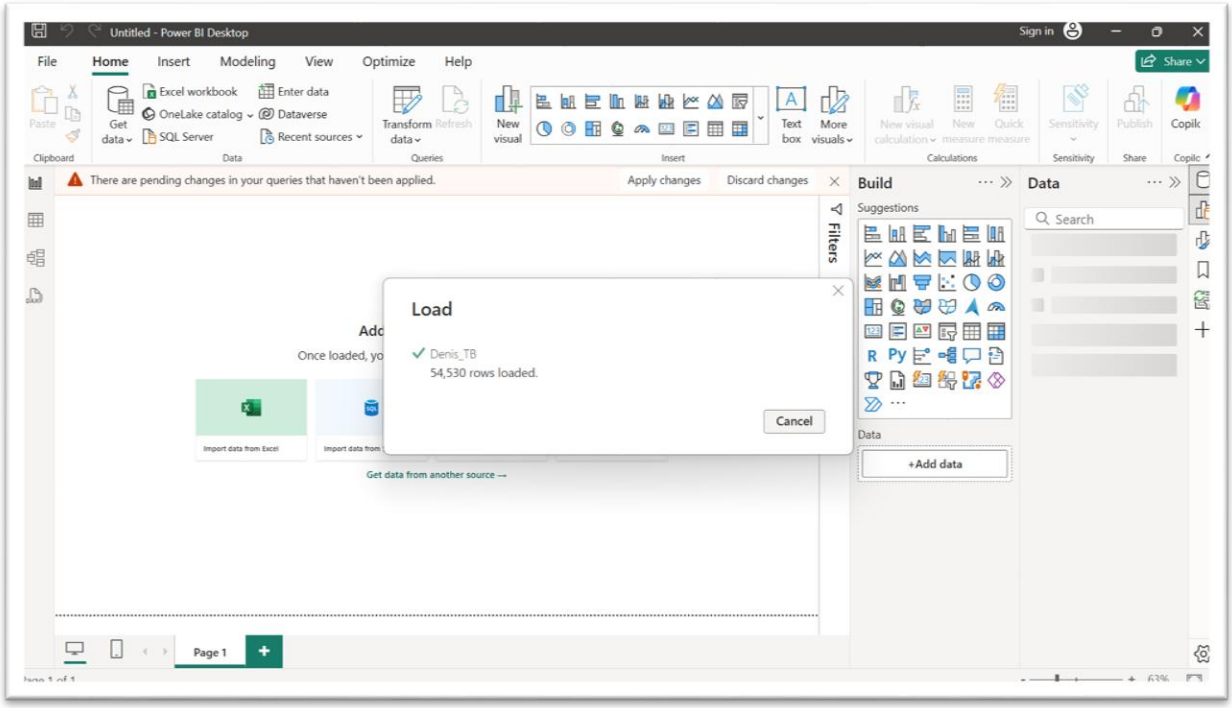


Shanti Jogi DE

Also data Moved to Azure SQL Database

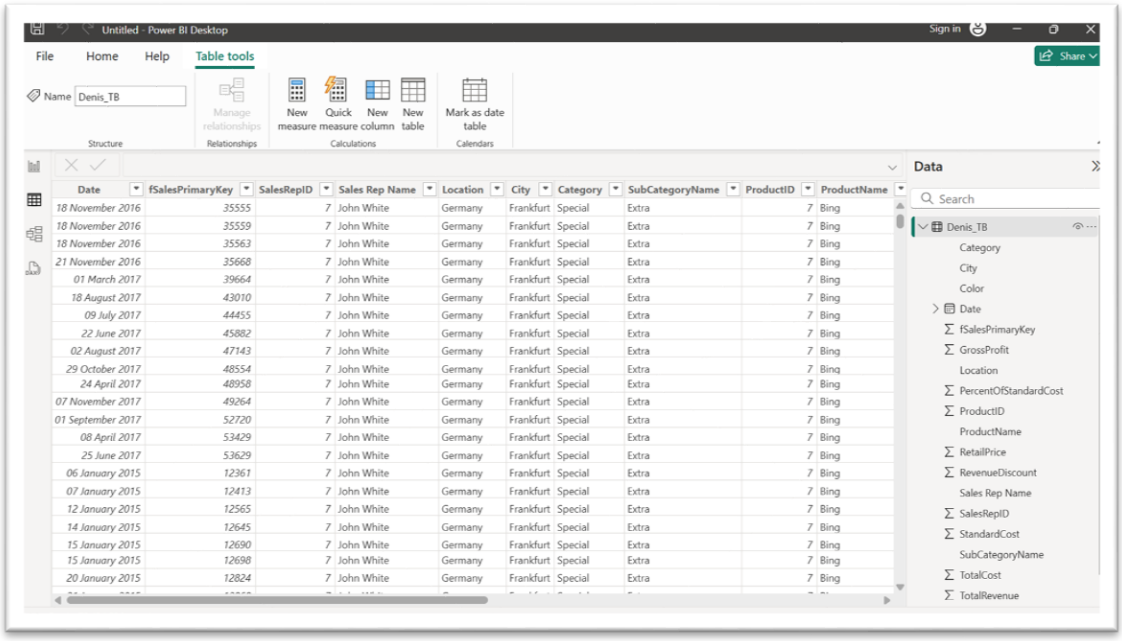


Connected this data to Power Bi vie Azure SQL Database Server



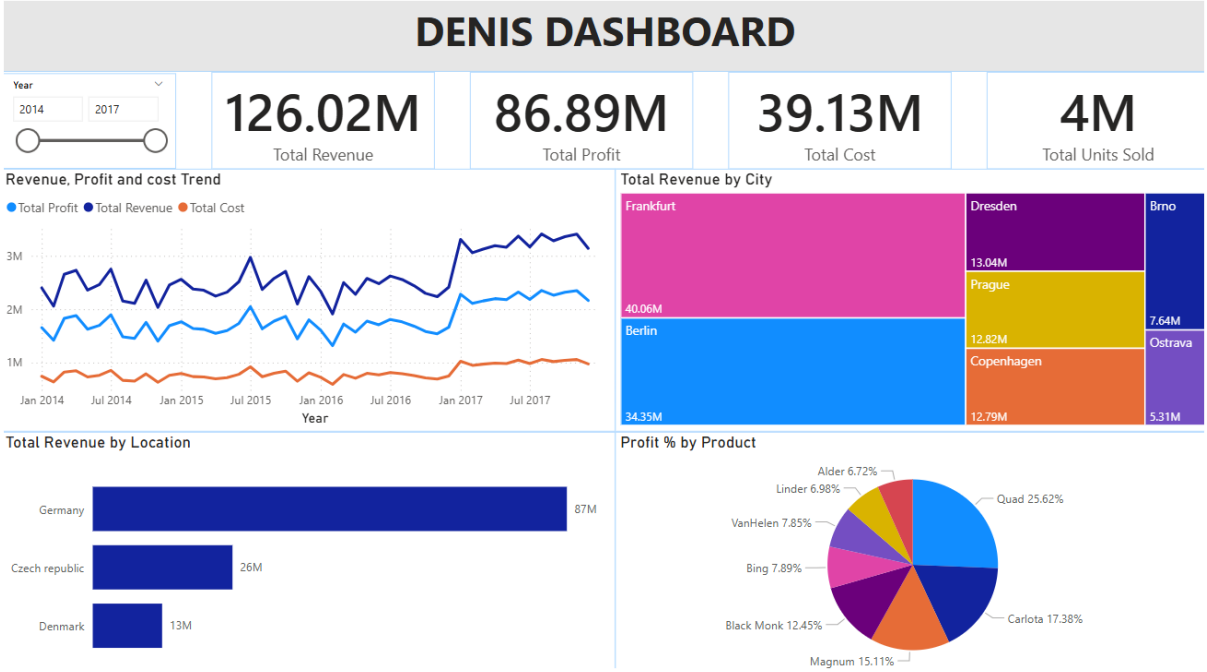
Shanti Jogi DE

Successful connected and loaded the data



Date	fSalesPrimaryKey	SalesRepID	Sales Rep Name	Location	City	Category	SubCategoryName	ProductID	ProductName
18 November 2016	35555	7	John White	Germany	Frankfurt	Special	Extra	7	Bing
18 November 2016	35559	7	John White	Germany	Frankfurt	Special	Extra	7	Bing
18 November 2016	35563	7	John White	Germany	Frankfurt	Special	Extra	7	Bing
21 November 2016	35668	7	John White	Germany	Frankfurt	Special	Extra	7	Bing
01 March 2017	39664	7	John White	Germany	Frankfurt	Special	Extra	7	Bing
18 August 2017	43010	7	John White	Germany	Frankfurt	Special	Extra	7	Bing
09 July 2017	44455	7	John White	Germany	Frankfurt	Special	Extra	7	Bing
22 June 2017	45882	7	John White	Germany	Frankfurt	Special	Extra	7	Bing
02 August 2017	47143	7	John White	Germany	Frankfurt	Special	Extra	7	Bing
29 October 2017	48554	7	John White	Germany	Frankfurt	Special	Extra	7	Bing
24 April 2017	48958	7	John White	Germany	Frankfurt	Special	Extra	7	Bing
07 November 2017	49264	7	John White	Germany	Frankfurt	Special	Extra	7	Bing
01 September 2017	52720	7	John White	Germany	Frankfurt	Special	Extra	7	Bing
08 April 2017	53429	7	John White	Germany	Frankfurt	Special	Extra	7	Bing
25 June 2017	53629	7	John White	Germany	Frankfurt	Special	Extra	7	Bing
06 January 2015	12361	7	John White	Germany	Frankfurt	Special	Extra	7	Bing
07 January 2015	12413	7	John White	Germany	Frankfurt	Special	Extra	7	Bing
12 January 2015	12565	7	John White	Germany	Frankfurt	Special	Extra	7	Bing
14 January 2015	12645	7	John White	Germany	Frankfurt	Special	Extra	7	Bing
15 January 2015	12690	7	John White	Germany	Frankfurt	Special	Extra	7	Bing
15 January 2015	12698	7	John White	Germany	Frankfurt	Special	Extra	7	Bing
20 January 2015	12824	7	John White	Germany	Frankfurt	Special	Extra	7	Bing

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.

Shanti Jogi DE