Centralized Data Warehouse system

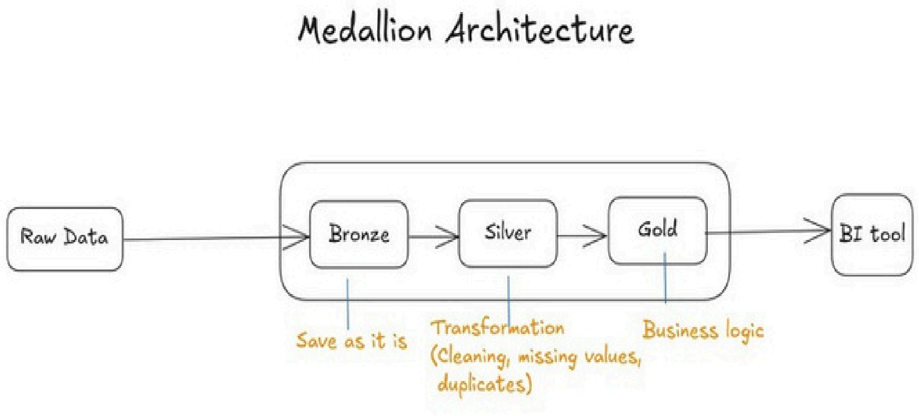
Introduction

The main goal of this project is to build a data in one place. This system will make it Centralized Data Warehouse system that brings together all sales-related Sales Insights for better reporting and business decisions.

Project Objective

The main goal of this project is to build a Centralized Data Repository that acts as a single source. All useful data — from SalesRep, Product, Sales info, geography, Categories, Subcategories — will be collected and managed through a well-defined pipeline. The data will be processed using the Medallion Architecture, which separates the data into three stages:

- Bronze Layer: Raw data is ingested with no changes.
- Silver Layer: Data is cleaned, checked, and linked together.
- Gold Layer: Final business-ready data is prepared for reporting and analysis.

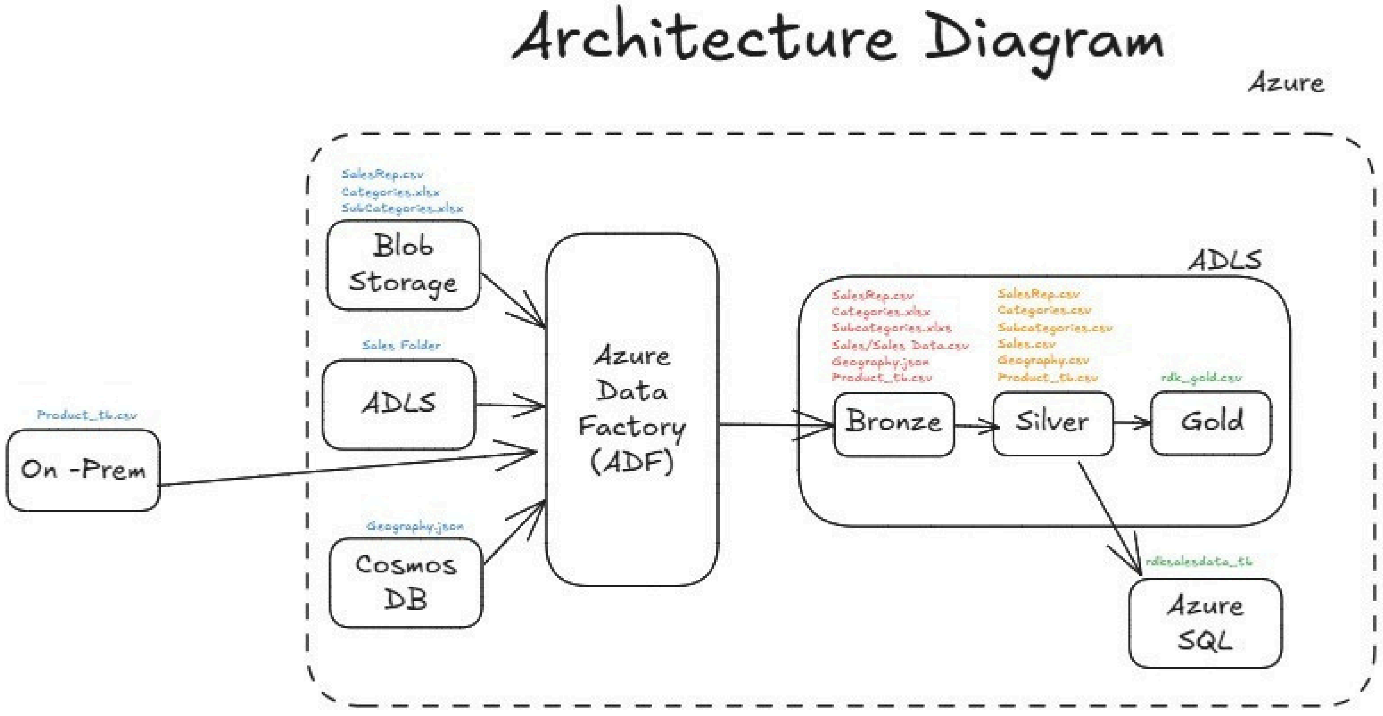


Layer	Purpose	Type of Data Processed
Bronze Layer	Raw data un processed files multiple sources	Original data (JSON, CSV, Excel) as-is.
Silver Layer	Cleans the data, applies schema validation, and builds relationships.	Cleaned and enriched data, with joins and filtering.
Gold Layer	Finalized, business-ready data for dashboards, reports, and Aggregated and curated datasets optimized	ML models. for analytics.

Data Ingestion & Processing Pipeline

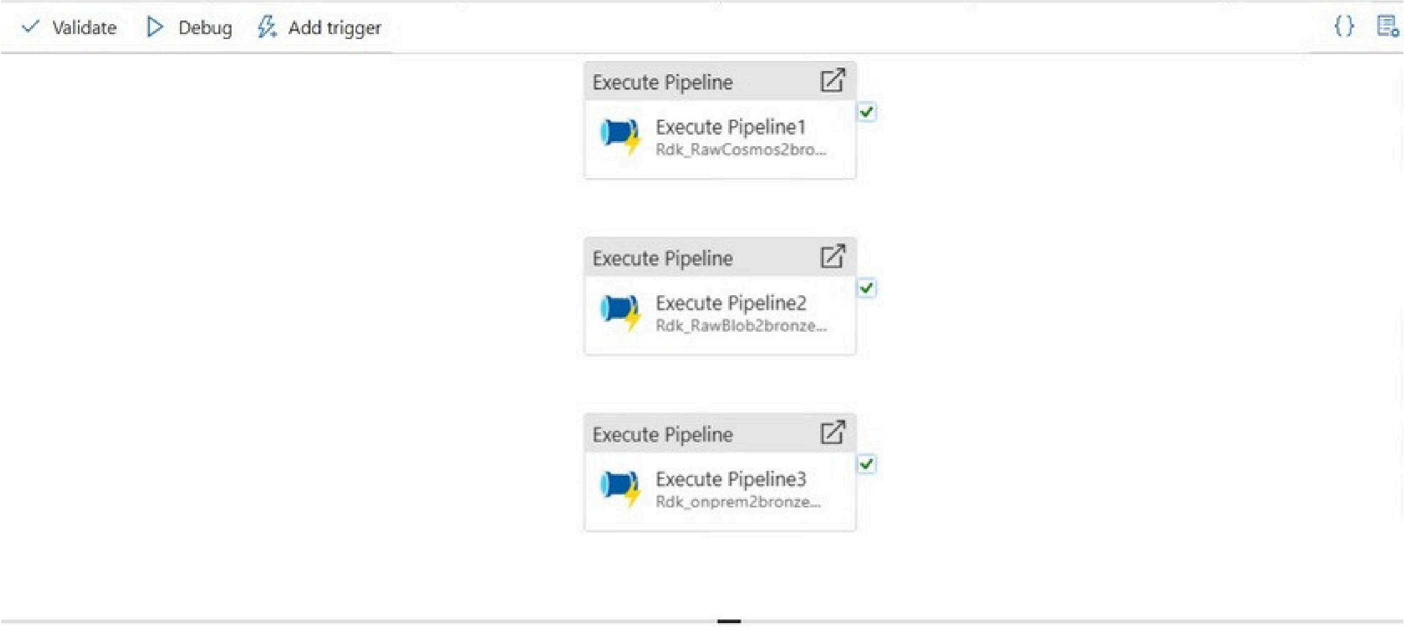
Step	Description
1. Data Ingestion	Load raw data from various formats: - JSON (Geography) - CSV (SalesRep, Product, Sales info, geography) - Excel (Categories, Subcategories)
2. Bronze Layer	Store the raw data as-is in the data lake for backup and traceability
3. Silver Layer	Clean the data (remove nulls, fix formats)
4. Gold Layer	Prepare final, optimized tables for business join related tables

Data Flow Diagram



Raw data to Bronze Layer

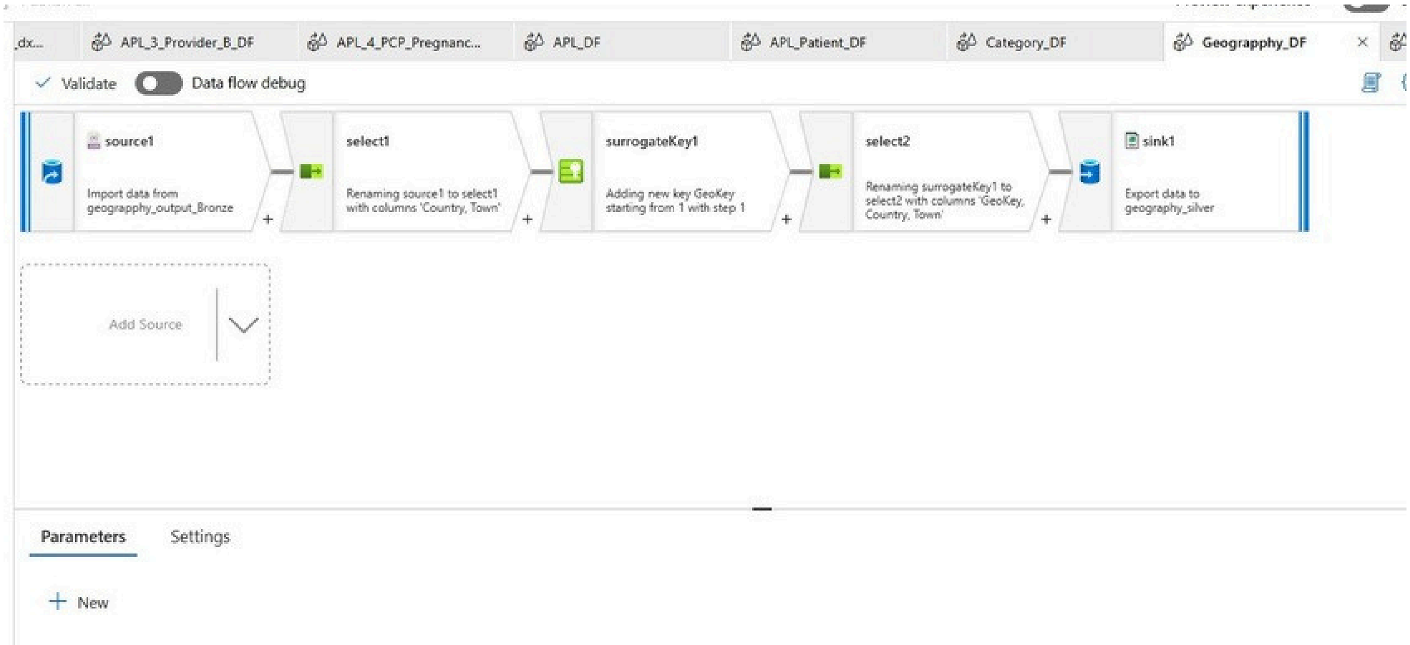
Raw data from different sources like blob storage, Azure data lake Gen 2(ADLS Gen 2), On – prem MySQL, Azure cosmos db for mongoDB moved to Azure data lake Gen 2 (ADLS) using Azure Data factory copy data activity. We have created input datasets as On – prem mysql , blob storage, ADLS and cosmosdb for mongo with link services and also created output dataset as ADLS with link services from azure data factory to ADLS. Based on this source and sink selected in copy activity. Created separate pipeline for all these and execute these pipelines one main pipeline.



Raw data to bronze main pipeline

Bronze to Silver Layer

Created Data flow and select file from bronze layer which we saved in ADLS storage. Select the source and check each column if it having duplicates or missing values, we will work on this and made changes on that. These dataflows called in a pipeline, here all dimension data's are succeeded only sales data moved to silver folder. And moved to ADLS storage in silver layer.



source1: Import data from geograpphy_output_Bronze

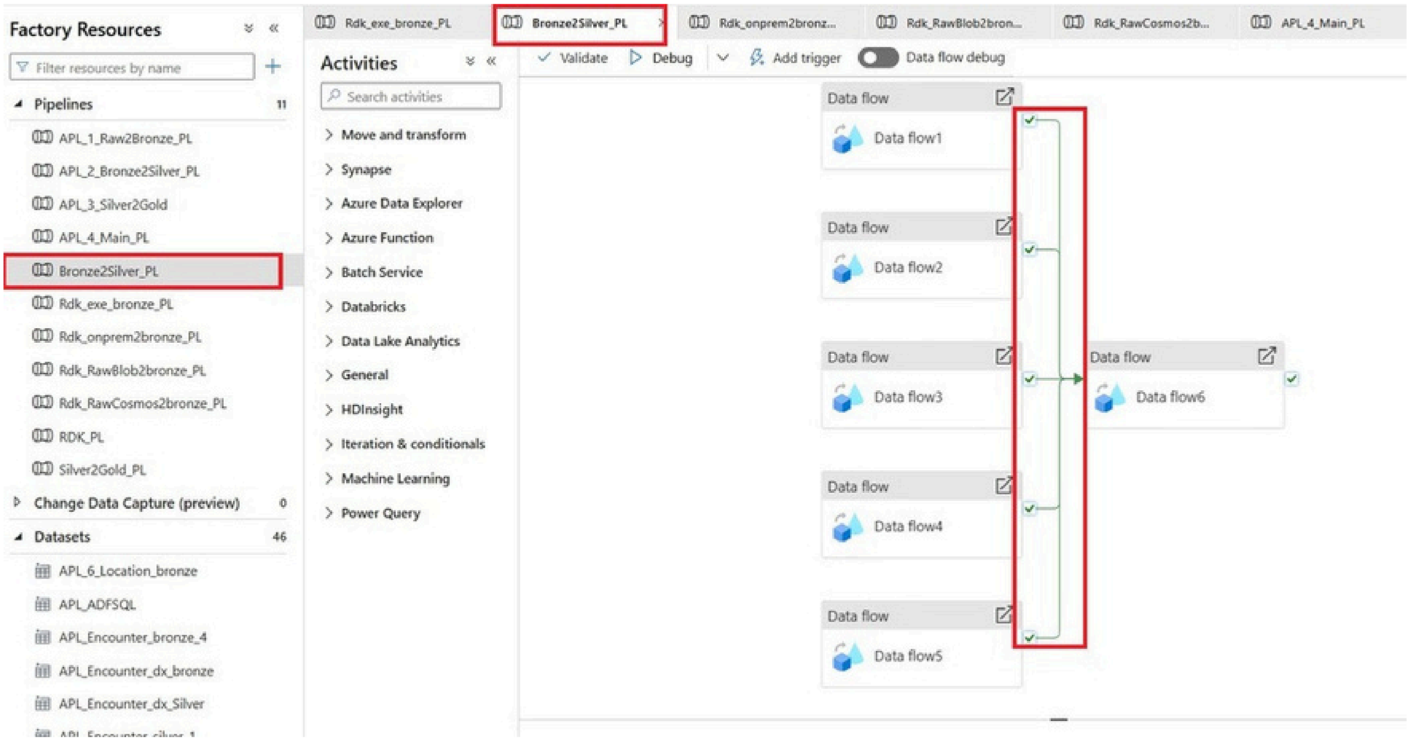
select1: Renaming source1 to select1 with columns 'Country, Town'

surrogateKey1: Adding new key GeoKey starting from 1 with step 1

select2: Renaming surrogateKey1 to select2 with columns 'GeoKey, Country, Town'

sink1: Export data to geography_silver

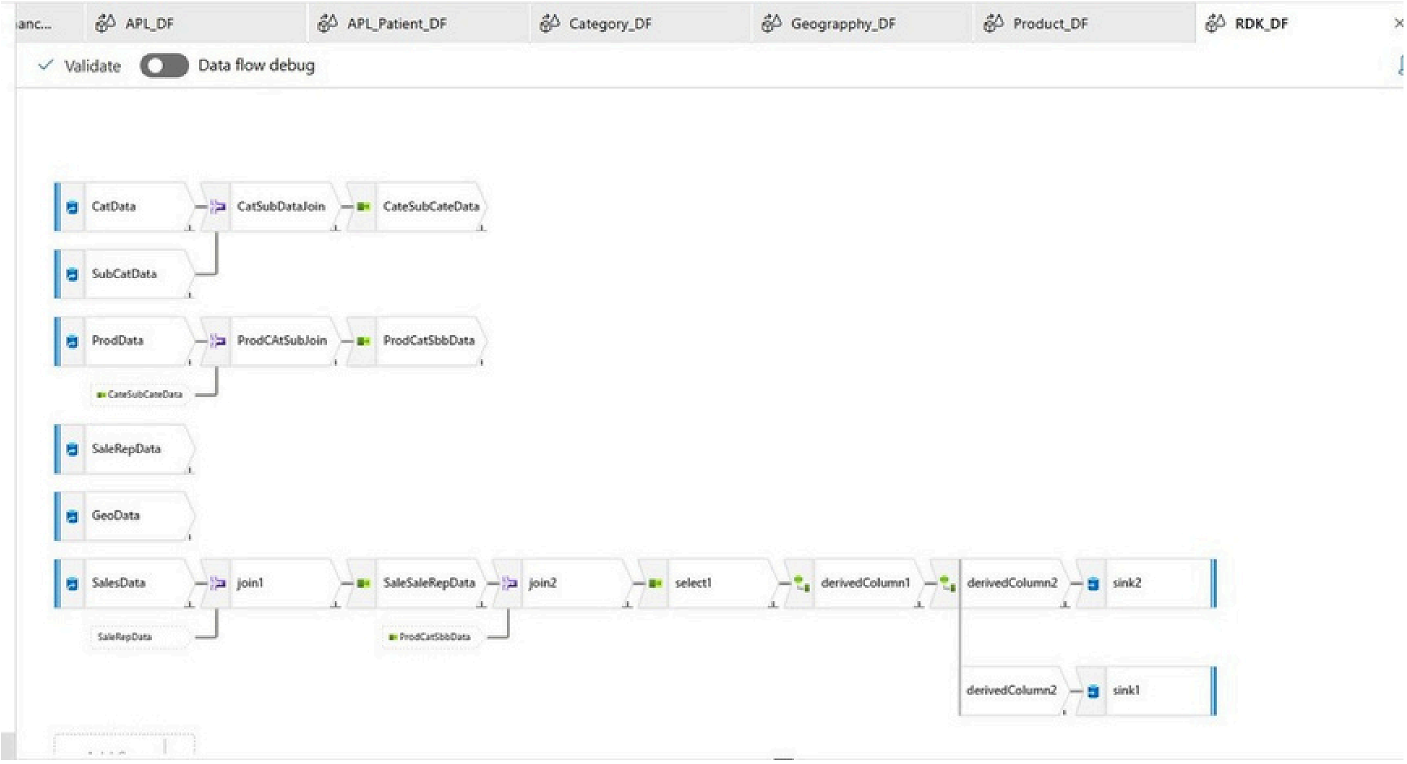
Bronze to Silver one Data flow - Geography



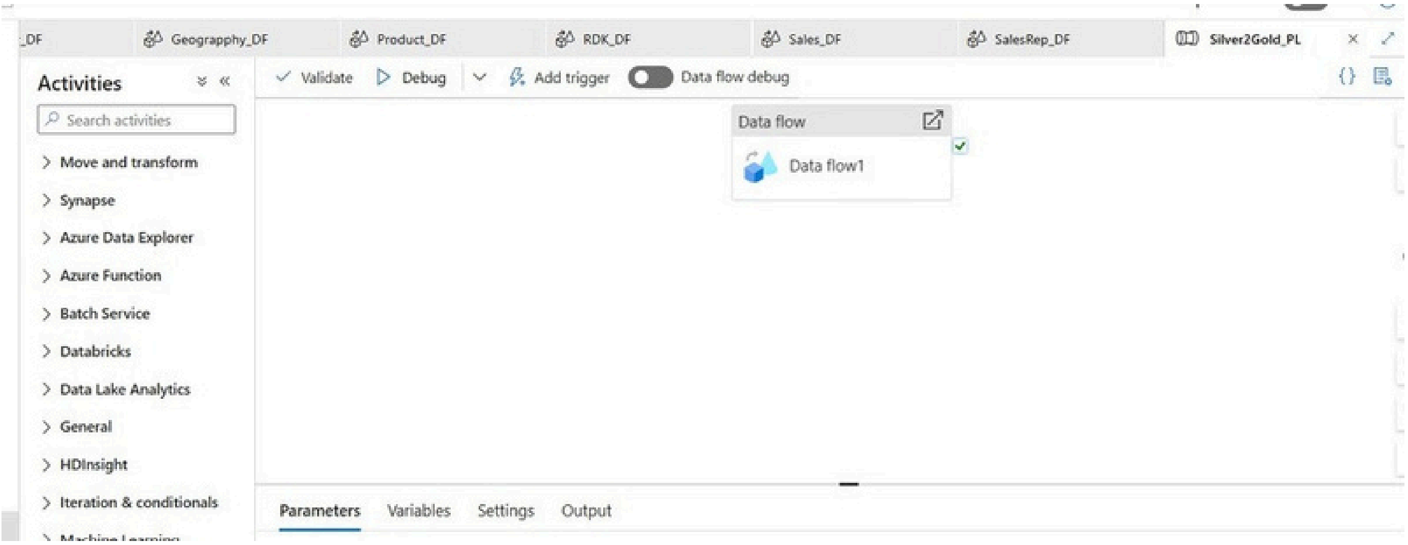
Bronze to Silver Pipeline

Silver to Gold Layer

Created Data flow and select file from silver layer which we saved in ADLS storage. Select the source and combine all data and consolidated in single file and it's moved to ADLS storage in gold layer and also moved to Azure SQL database based on selecting sink as azure SQL.

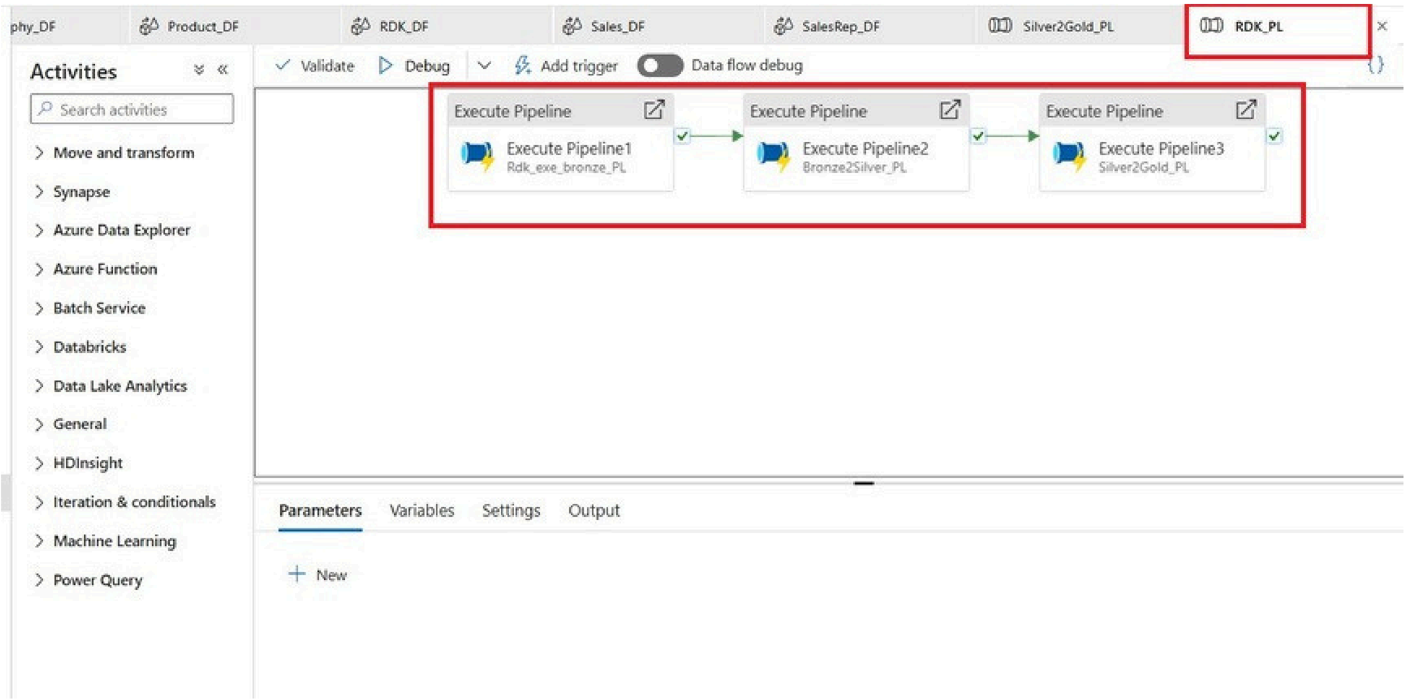


Gold data flow

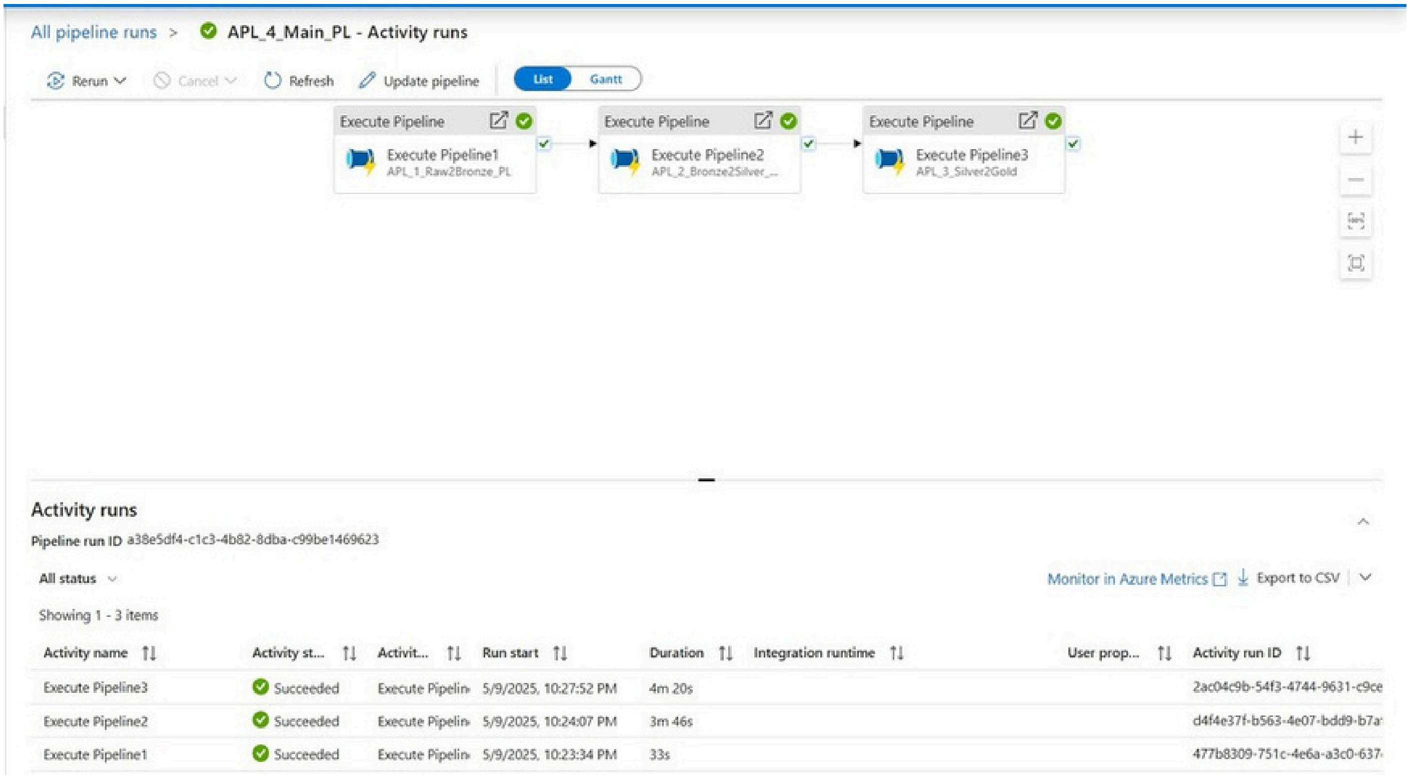


Silver to Gold Pipeline

For automatically run row to gold process, we created a main pipeline it having all these above pipelines using execute pipeline.



Main Pipeline



All pipeline runs > APL_4_Main_PL - Activity runs

Activity runs

Pipeline run ID a38e5df4-c1c3-4b82-8dba-c99be1469623

All status

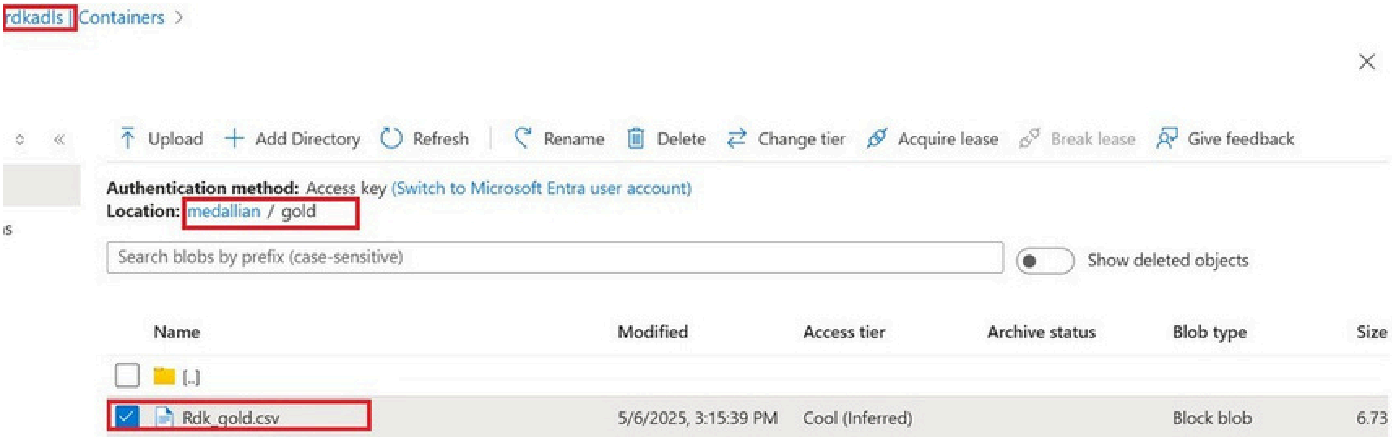
Monitor in Azure Metrics | Export to CSV

Showing 1 - 3 items

Activity name	Activity st...	Activit...	Run start	Duration	Integration runtime	User prop...	Activity run ID
Execute Pipeline3	✓ Succeeded	Execute Pipelin	5/9/2025, 10:27:52 PM	4m 20s			2ac04c9b-54f3-4744-9631-c9ce
Execute Pipeline2	✓ Succeeded	Execute Pipelin	5/9/2025, 10:24:07 PM	3m 46s			d4f4e37f-b563-4e07-bdd9-b7a
Execute Pipeline1	✓ Succeeded	Execute Pipelin	5/9/2025, 10:23:34 PM	33s			477b8309-751c-4e6a-a3c0-637

Succeeded main pipeline -status

Summary



rdkadls | Containers >

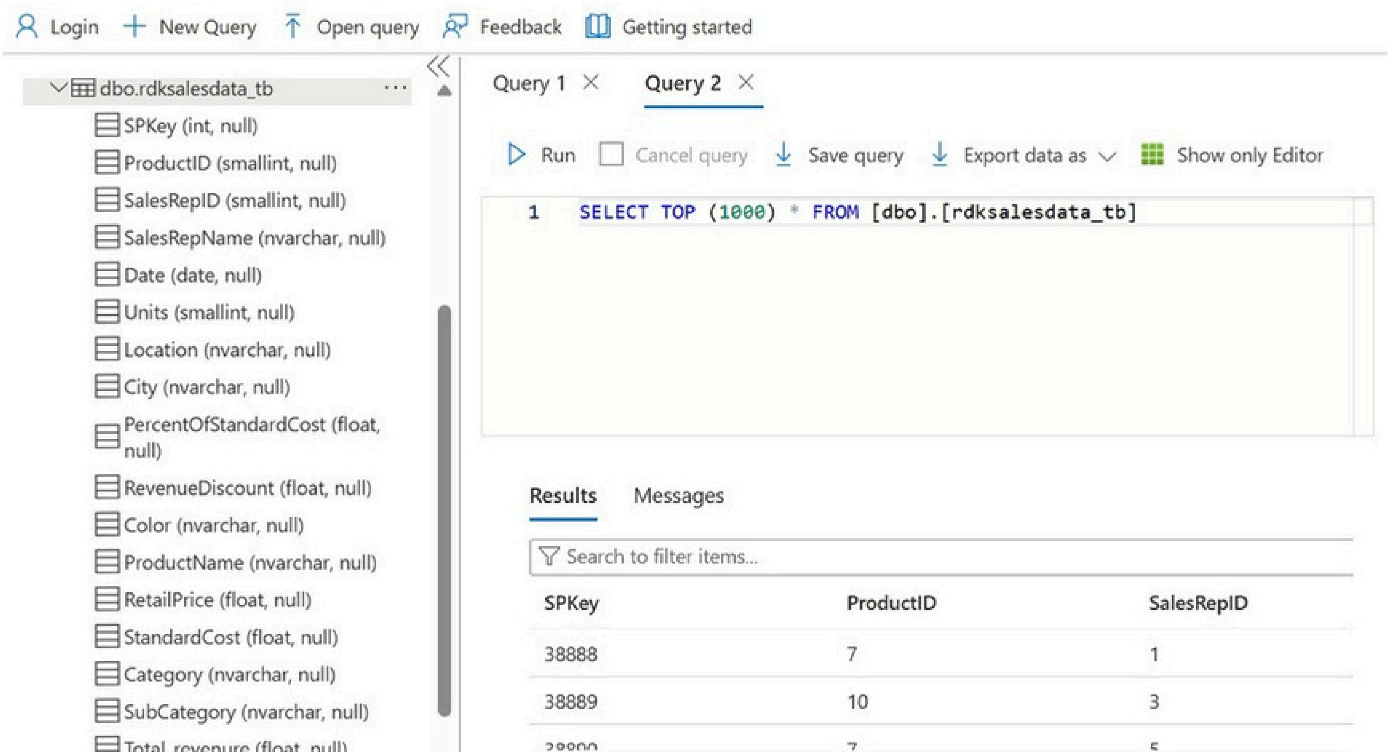
Authentication method: Access key (Switch to Microsoft Entra user account)
Location: medallian / gold

Search blobs by prefix (case-sensitive)

○ Show deleted objects

	Name	Modified	Access tier	Archive status	Blob type	Size
□	[..]					
☑	Rdk_gold.csv	5/6/2025, 3:15:39 PM	Cool (Inferred)		Block blob	6.73

Output data in Gold layer – ADLS gold folder.



dbo.rdksalesdata_tb
- SPKey (int, null)
- ProductID (smallint, null)
- SalesRepID (smallint, null)
- SalesRepName (nvarchar, null)
- Date (date, null)
- Units (smallint, null)
- Location (nvarchar, null)
- City (nvarchar, null)
- PercentOfStandardCost (float, null)
- RevenueDiscount (float, null)
- Color (nvarchar, null)
- ProductName (nvarchar, null)
- RetailPrice (float, null)
- StandardCost (float, null)
- Category (nvarchar, null)
- SubCategory (nvarchar, null)
- Total_revenure (float, null)

```
SELECT TOP (1000) * FROM [dbo].[rdksalesdata_tb]
```

SPKey	ProductID	SalesRepID
38888	7	1
38889	10	3
38890	7	5

Consolidated data in Azure SQL database table.

Centralized Data Warehouse system implemented an automated, scalable pipeline using Azure Data Factory,

aligned with Medallion Architecture (Raw to Bronze, Bronze to Silver and Silver to Gold layer). It met all data ingestion, transformation, and reporting needs, delivering a refined gold layer optimized single csv file with all necessary data.