# End-To-End Data Engineering Project using Microsoft Azure Data Factory and Azure Synapse Analytics

# Business Problem

## Context:

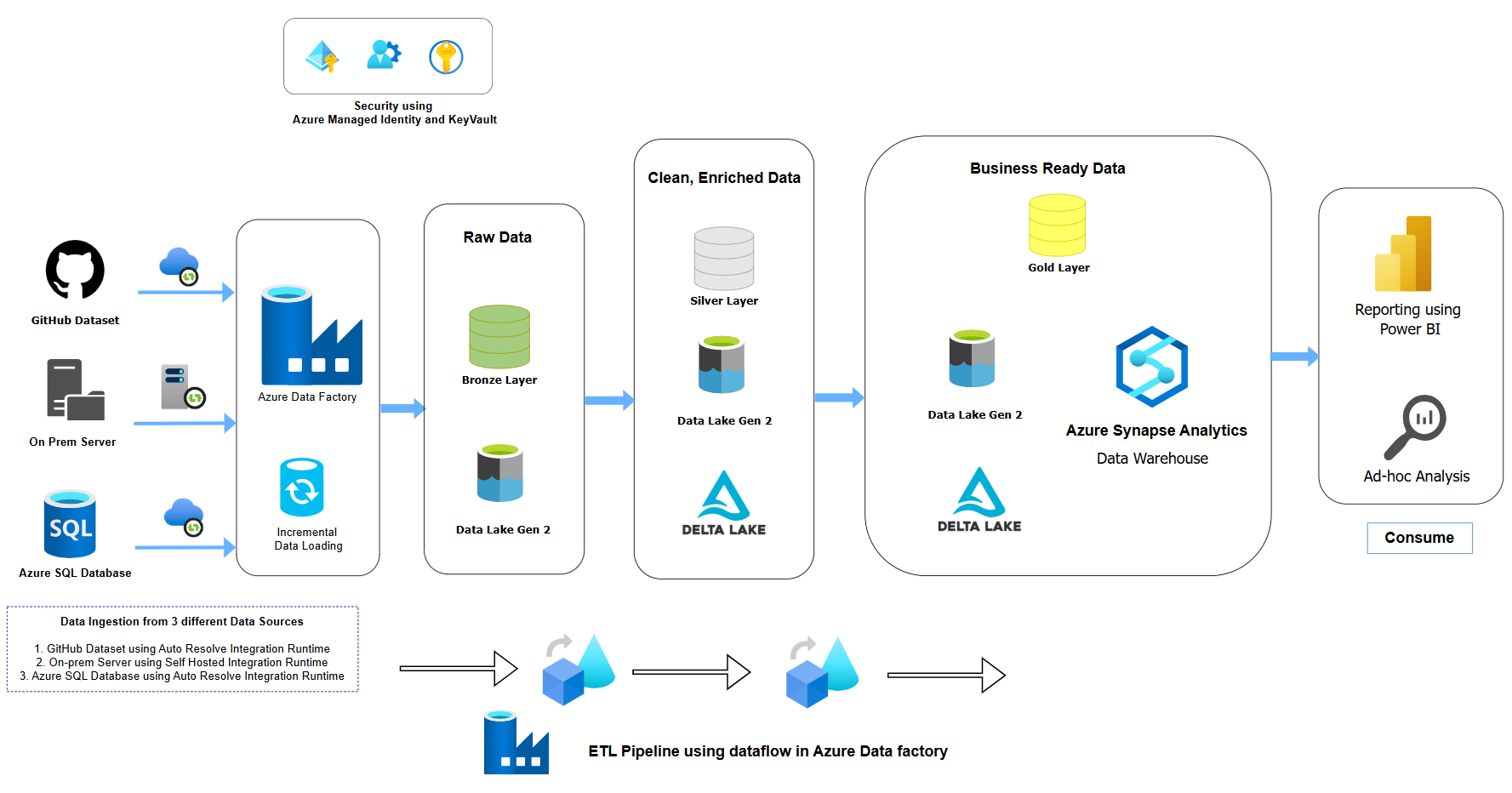
A retail company (or a similar enterprise in this case study) operates in multiple regions with diverse product lines and a network of resellers. Sales are driven by a distributed salesforce, and the company sets quarterly/annual sales targets at different levels (by product, region, or salesperson.

## Key Business Problems:

* No unified view of sales across category, brand, salesperson.
* Difficulty in tracking sales performance vs. targets
* Inefficient allocation of resources to underperforming areas
* No end-to-end data pipeline to process the data.

## Proposed Data-Driven Solution:

* Ingested raw CSV datasets into **Azure Data Lake Gen2** using **Azure Data Factory**
* Stored raw, untransformed data in the **Bronze Layer** in Azure Data Lake
* Cleaned, standardized, and enriched data using **ADF Data Flows** in the **Silver Layer**
* Applied business logic and store data in Delta format in **Gold Layer.** Build a Data Warehouse Solution in **Azure Synapse Analytics**
* Secured the pipeline with **Azure Managed Identity & Key Vault**
* Delivered analytics-ready data for visualization in **Power BI**



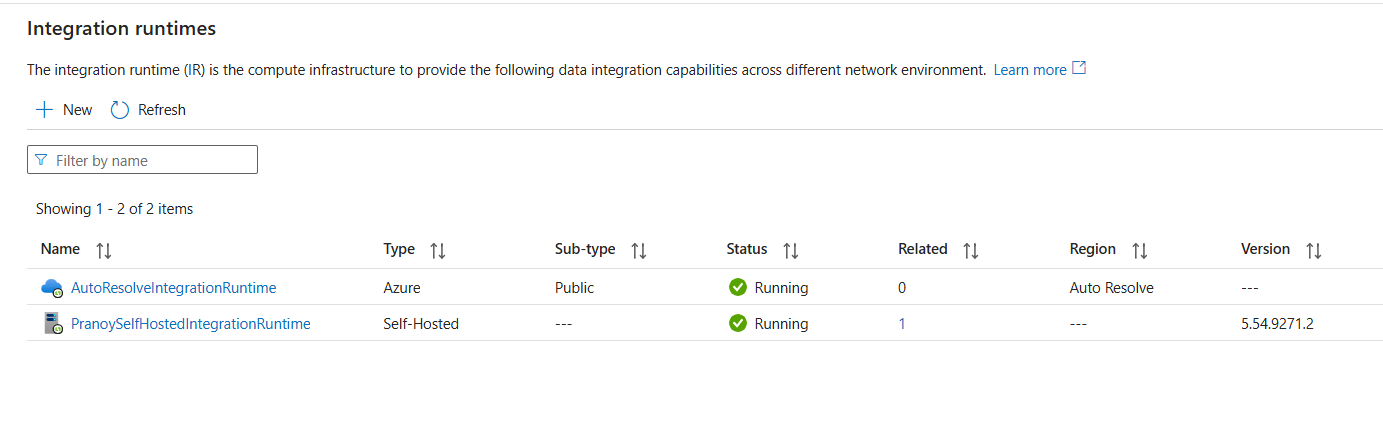
# Data Ingestion

Data have been stored in 3 different data sources.

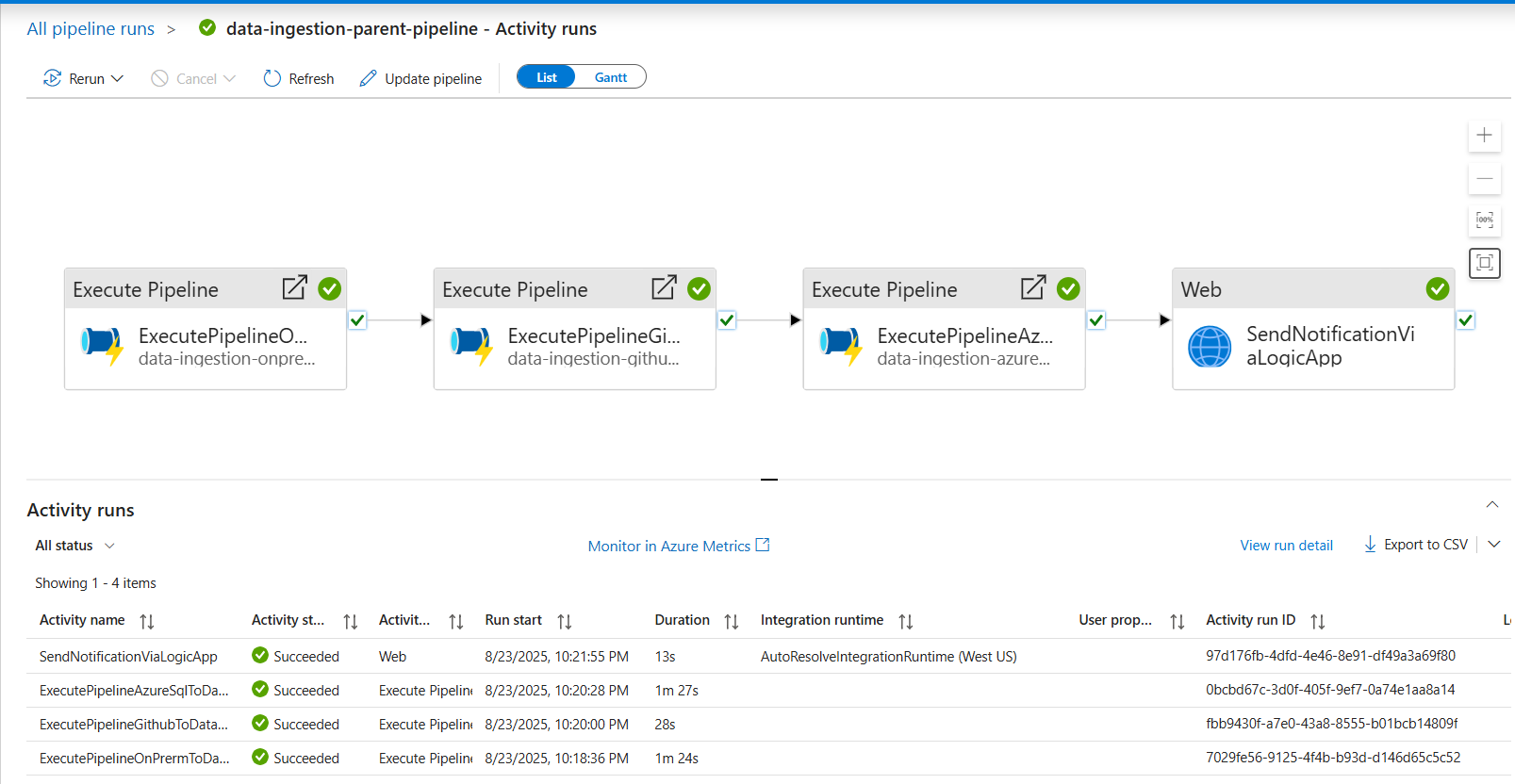
1. GitHub Storage in CSV format
2. On-Prem Server in CSV format
3. Azure SQL in structured format

Here I have used two different Integration Runtime:

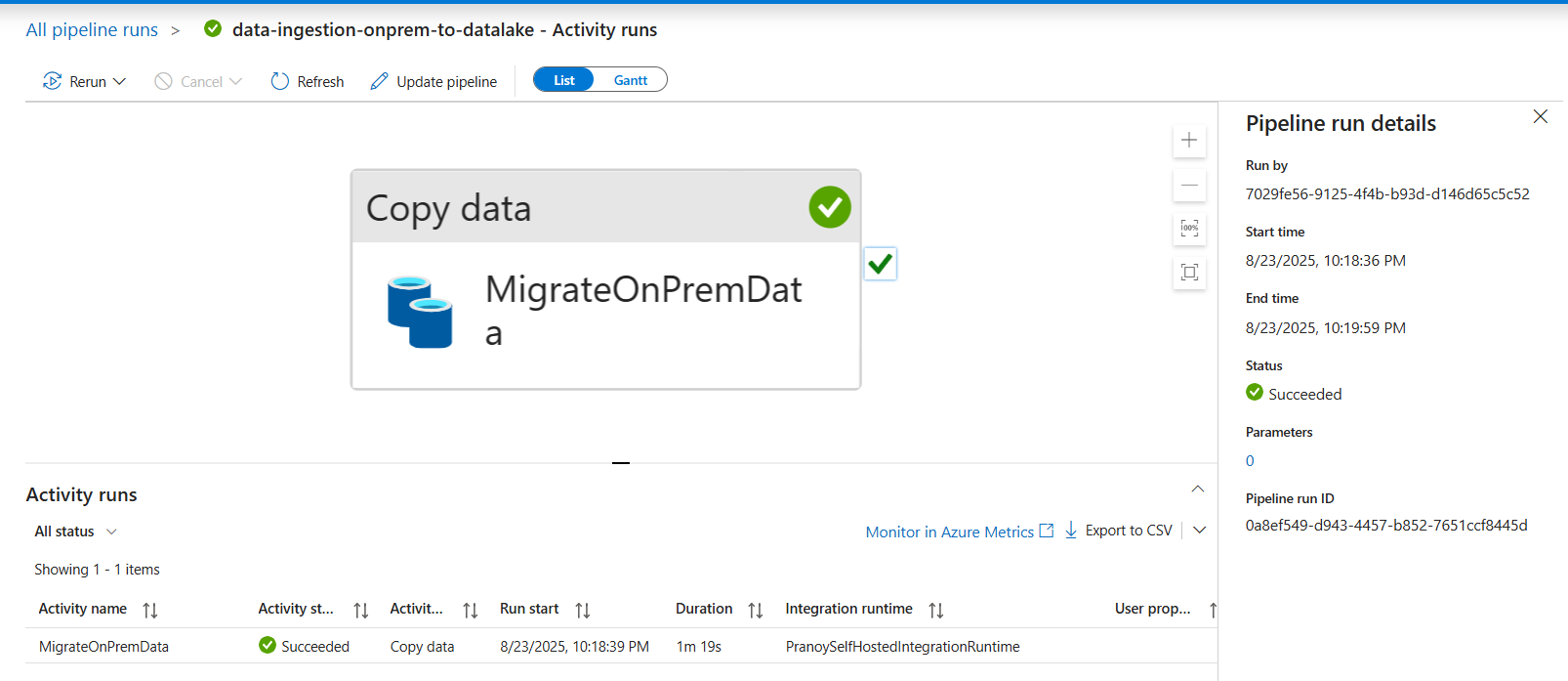
1. Self-Hosted Integration Runtime for migrating data from on-prem to cloud
2. Auto Resolve Integration Runtime for any other data operation



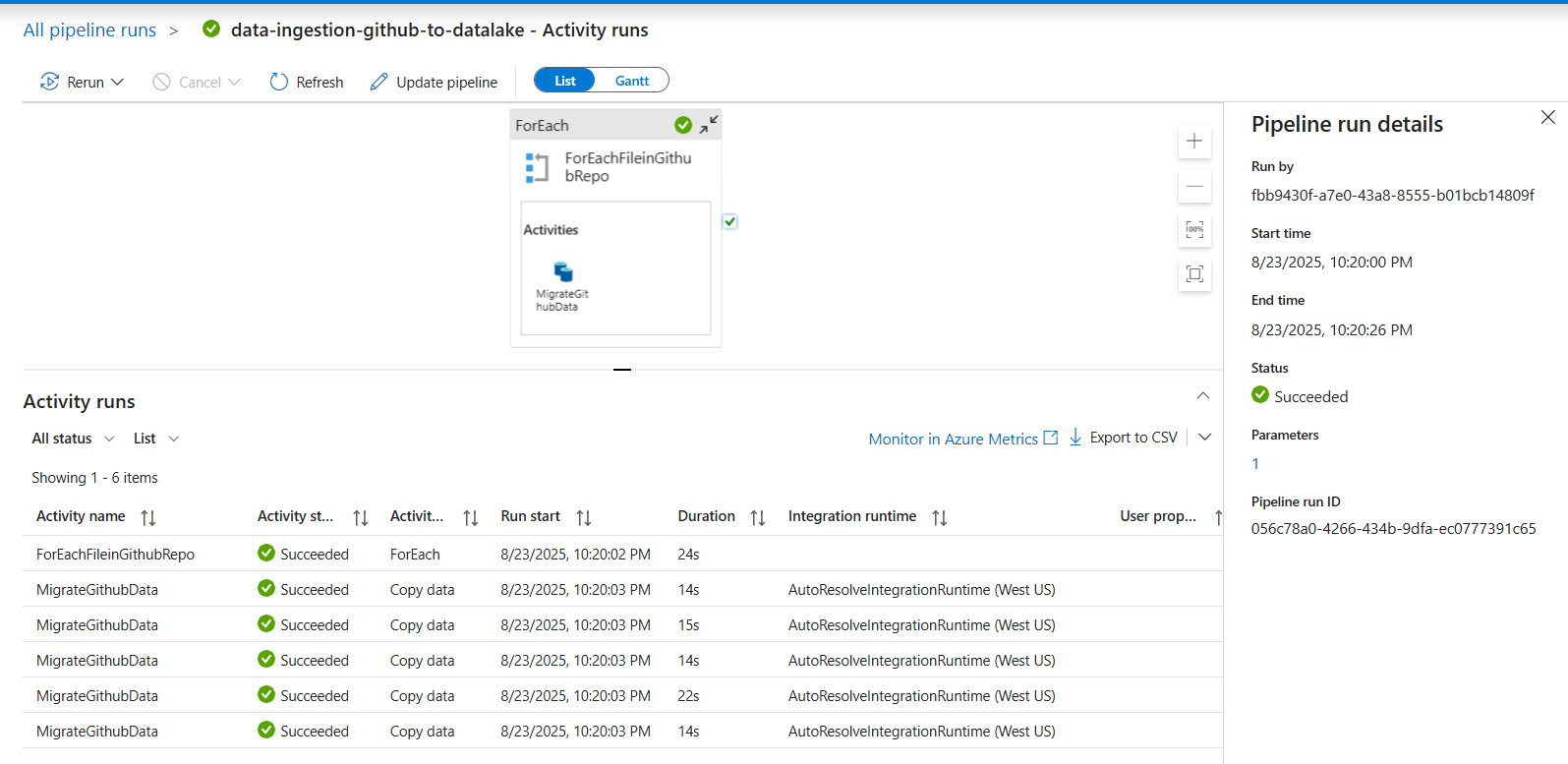
## Parent Pipeline Run



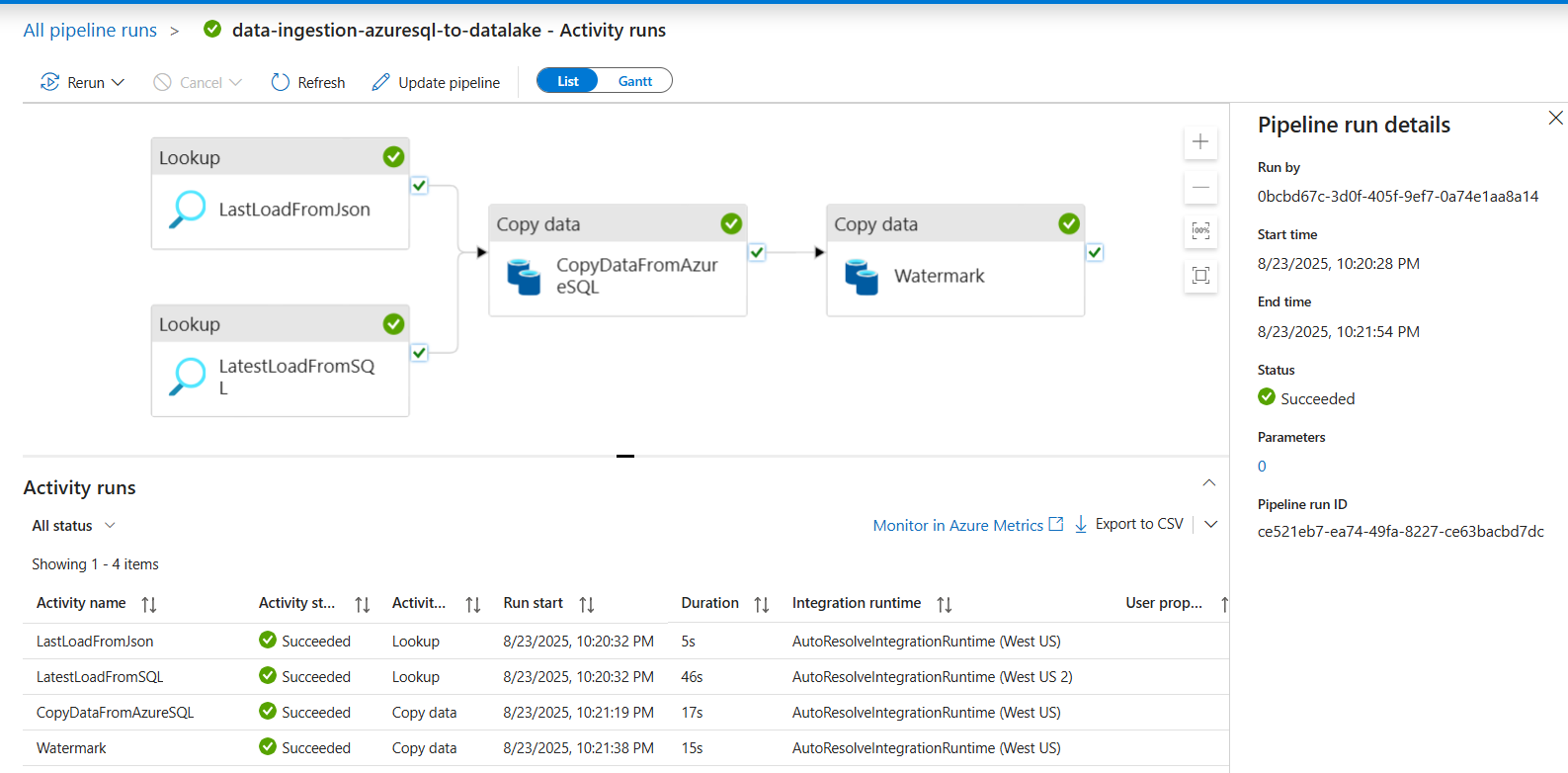
## Migrating Data from On-Prem to Azure Data Lake



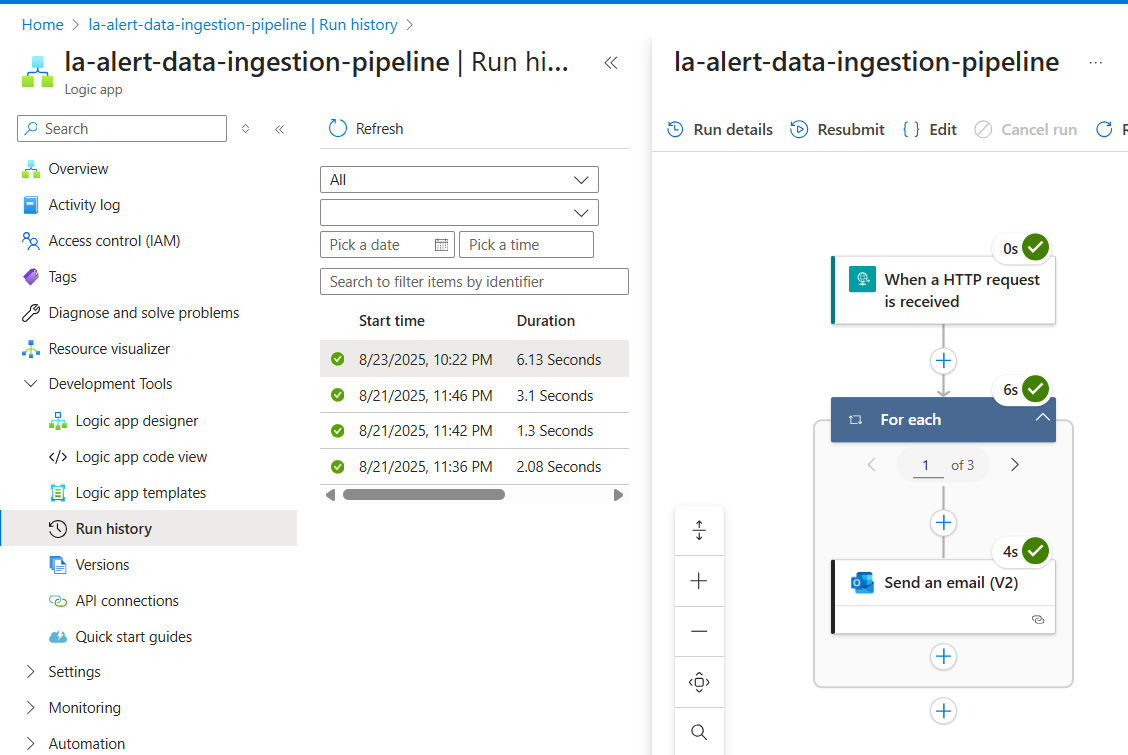
## Migrating Data from GitHub to Azure Data Lake



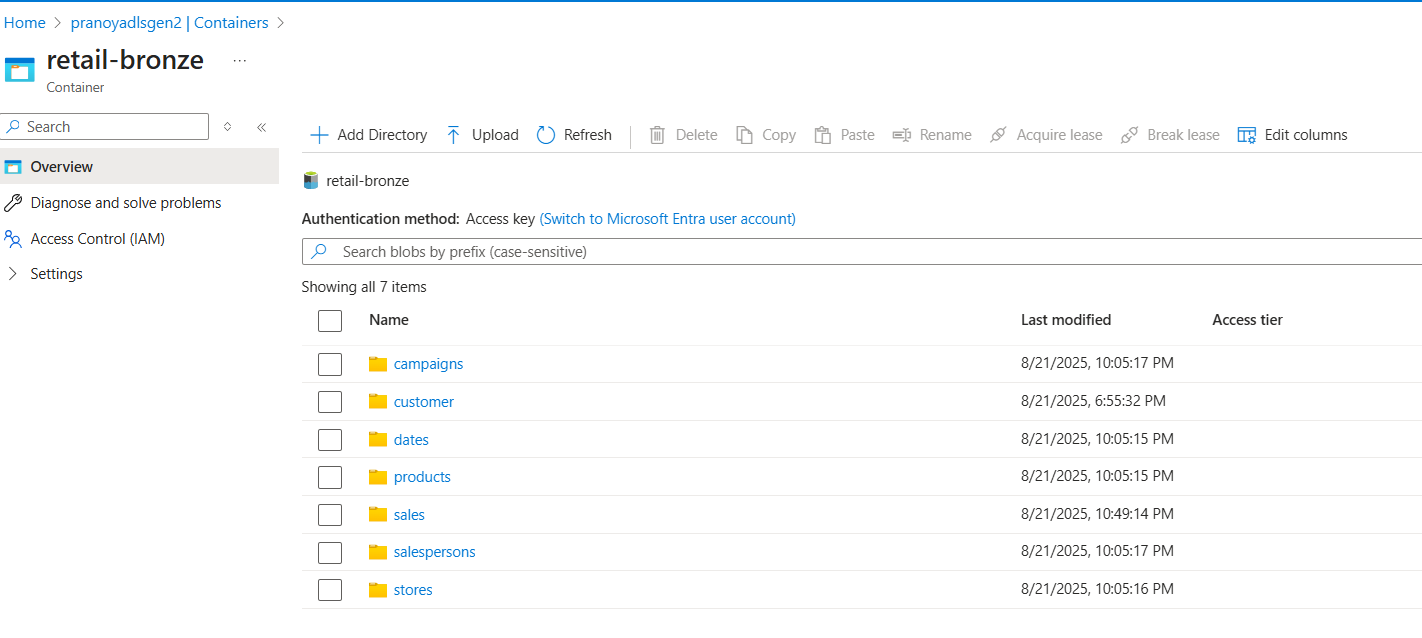
## Migrating Data from Azure SQL to Azure Data Lake + Incremental Data Loading



## Alert Notification using Azure Logic App



## Raw Data stored in Bronze Layer in Azure Data Lake (CSV format)

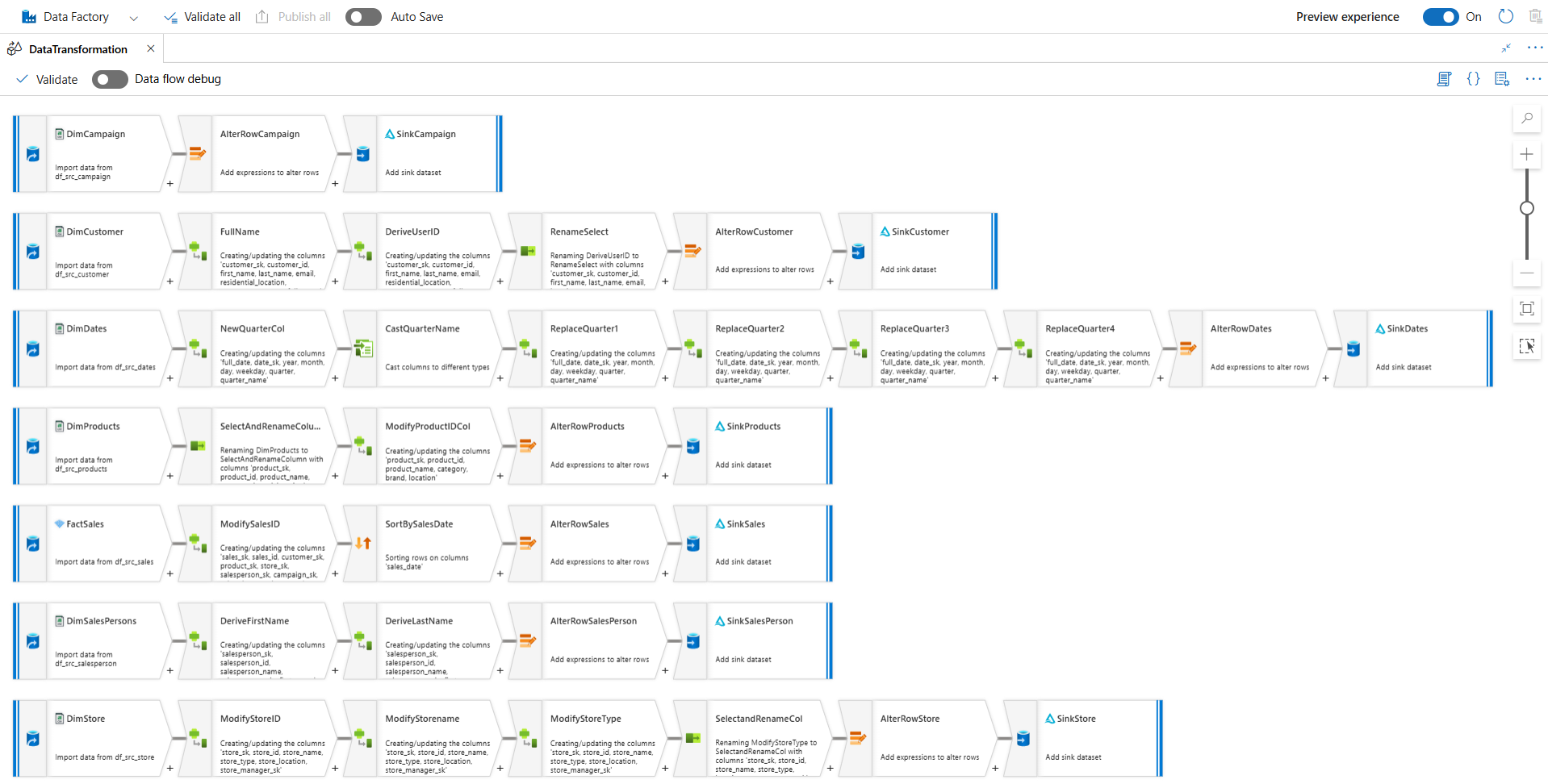


# Data Transformation

Data Transformation has been performed by using Data flow in Azure Data Factory. The data from **Bronze layer** has been extracted, cleaned, transformed and stored in **Silver Layer** in Delta Format.

1. ✅ **ACID Transactions on Data Lakes**
   * Delta brings **Atomicity, Consistency, Isolation, Durability** (ACID) to data lakes.
   * Prevents issues like partial writes, duplicate data, or corrupted files common with CSV/Parquet.
2. ✅ **Schema Enforcement & Evolution**
   * Enforces column types and structures to avoid “bad data” landing in production.
   * Supports **schema evolution** when new columns are added — without breaking existing pipelines.
3. ✅ **Time Travel (Data Versioning)**
   * Every write creates a new version of the table.
   * You can query older snapshots for **debugging, audits, or rollback** (e.g., “see data as of yesterday”).
4. ✅ **High Performance with Indexing & Caching**
   * Delta automatically creates metadata (transaction logs, indexes).
   * Faster queries compared to plain Parquet because it avoids scanning entire datasets.
5. ✅ **Unified Batch + Streaming**
   * Same Delta table can serve **streaming and batch pipelines** without separate architecture.
   * Great for real-time analytics and incremental data processing.

**Delta makes data lake reliable, consistent, and production-ready**, solving the weaknesses of raw CSV/Parquet.

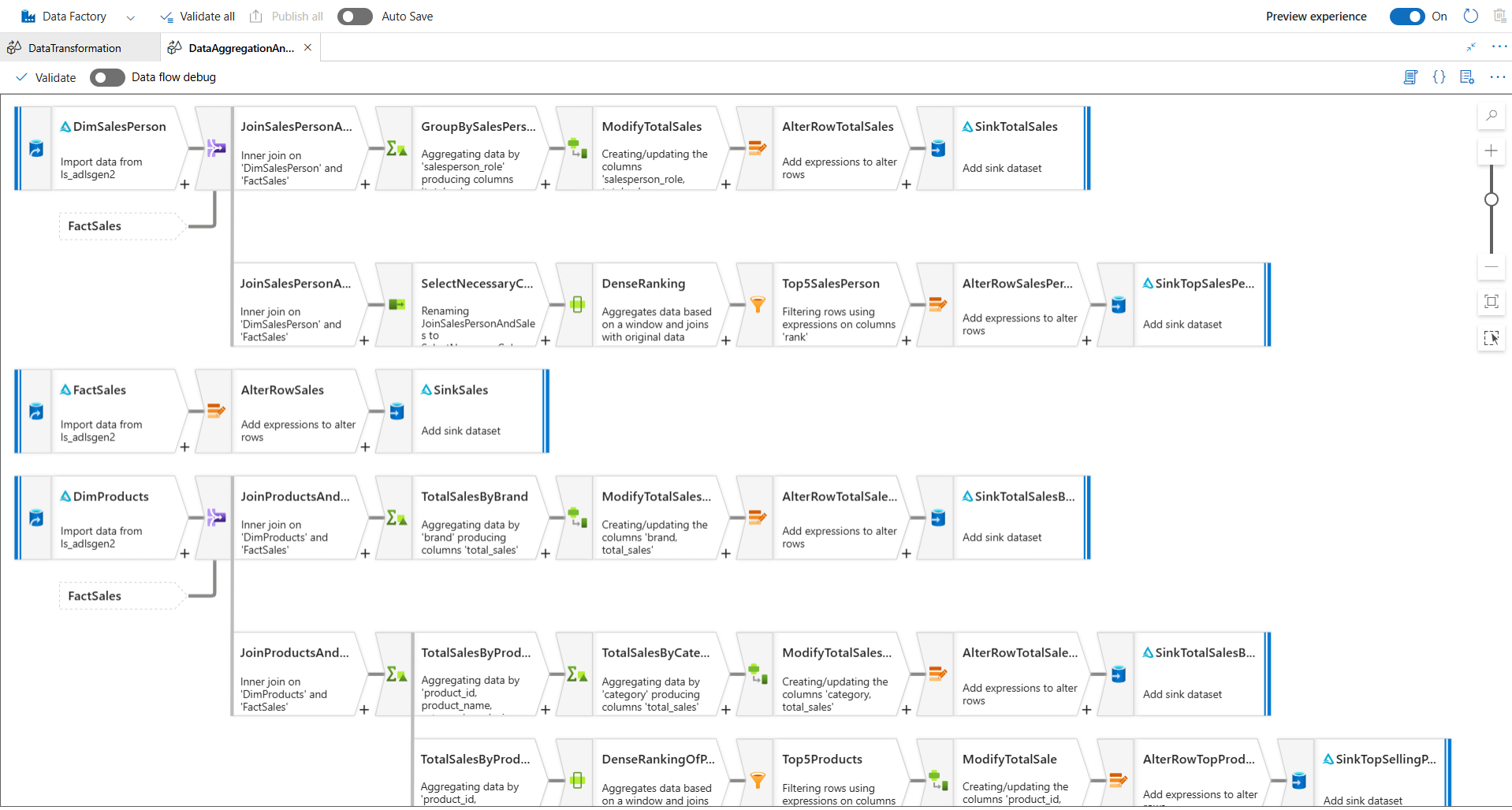


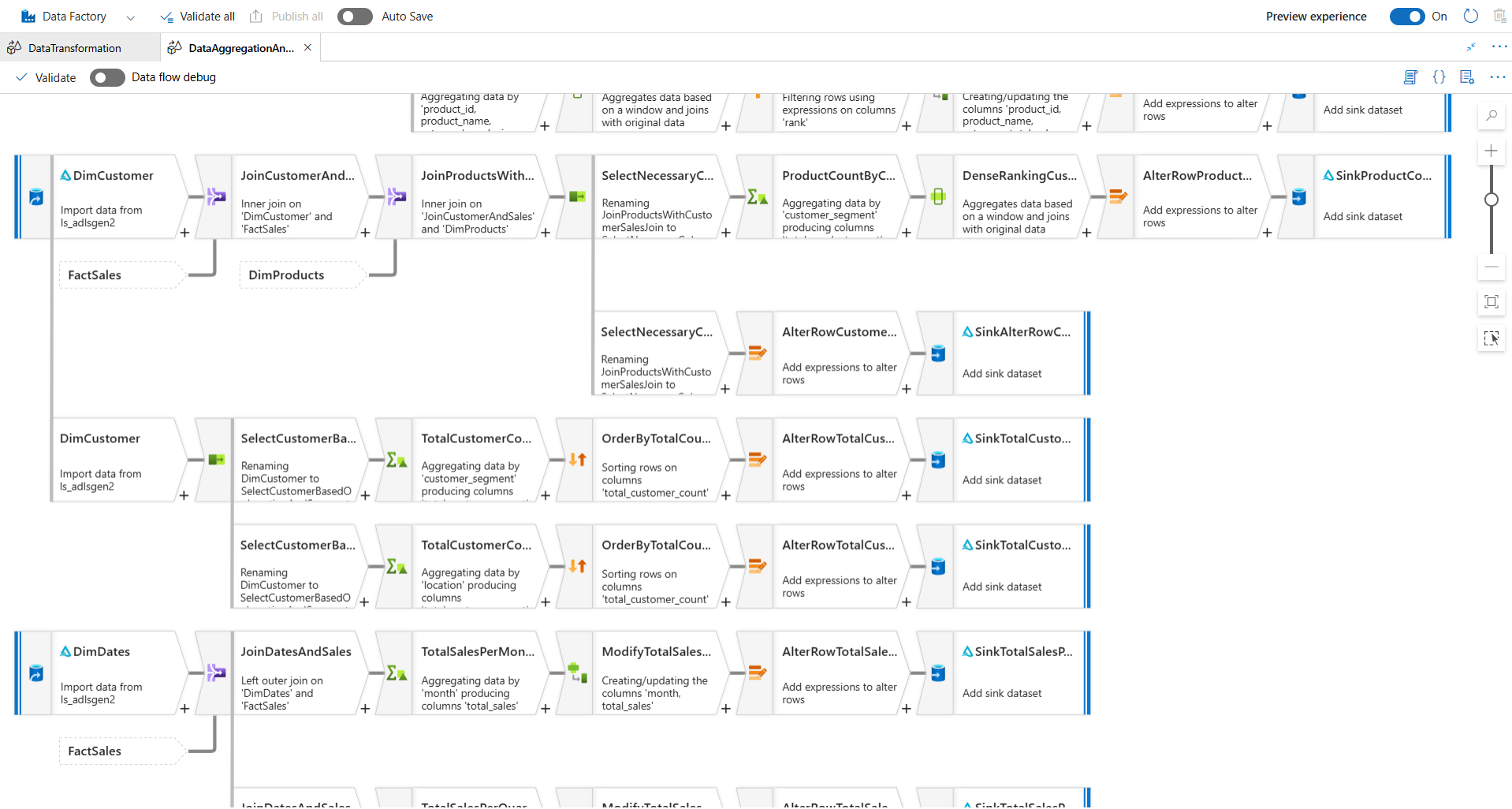
# Data Aggregation and Serving

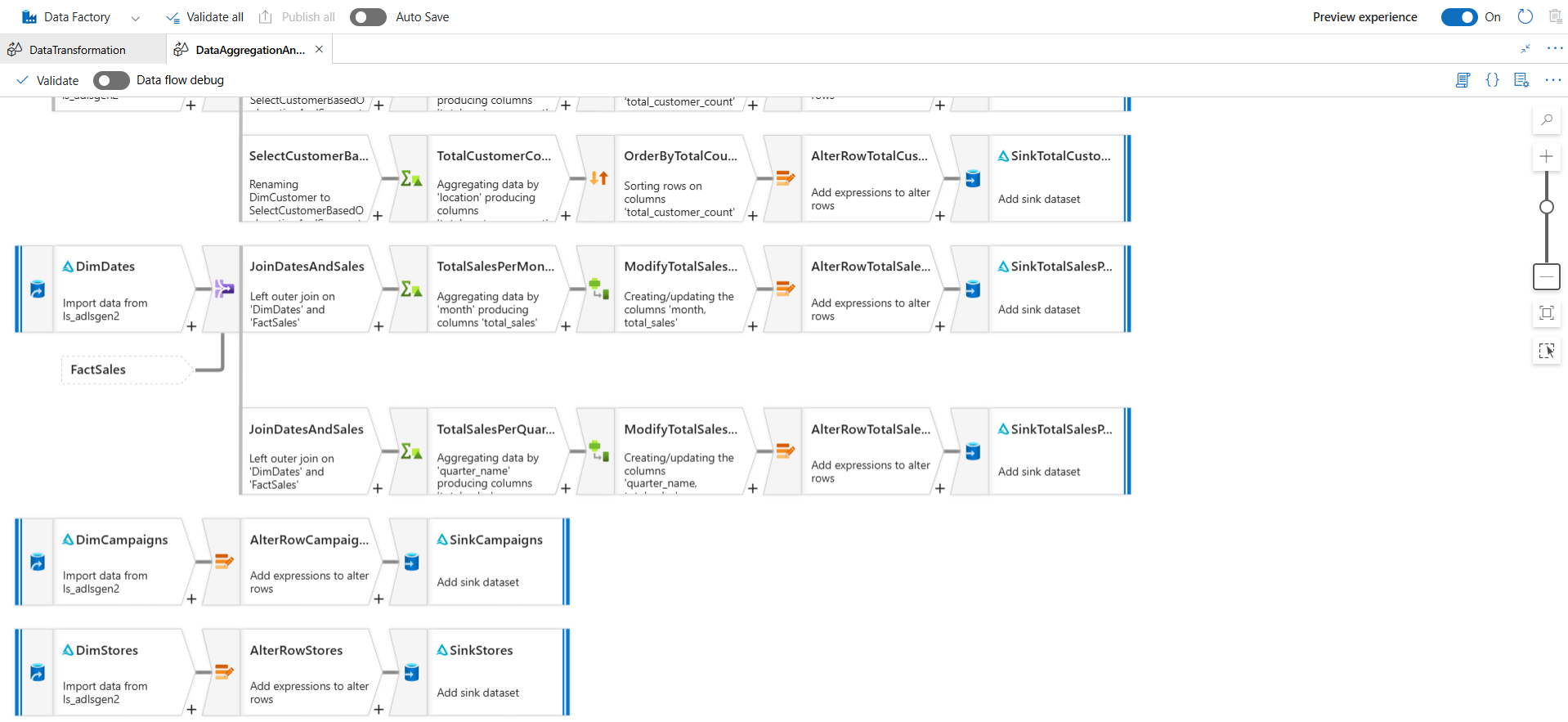
The **Gold Layer** represents the final curated and business-ready data model designed for reporting, dashboarding, and advanced analytics. Data from the **Silver Layer** is aggregated, enriched, and structured into fact and dimension models that directly support business KPIs and decision-making.

**Key Features of the Gold Layer:**

* ✅ **Business-Ready Data** – Provides clean, transformed, and aggregated datasets that can be directly consumed by BI tools like Power BI or Tableau.
* ✅ **Fact & Dimension Modeling** – Data is shaped into fact tables (sales, revenue, transactions) and dimension tables (customers, products, stores, regions) for intuitive analysis.
* ✅ **Advanced Transformations** – Includes **joins, ranking, grouping, and window functions** to derive insights such as *top-performing salespersons, best-selling products, or customer churn indicators*.
* ✅ **KPI Calculation** – Pre-computed measures (e.g., total sales, top products etc.) speed up dashboard performance and ensure consistency across analytics.
* ✅ **Actionable Insights** – Enables stakeholders to monitor trends, anomalies, and performance metrics that directly guide strategic decisions.

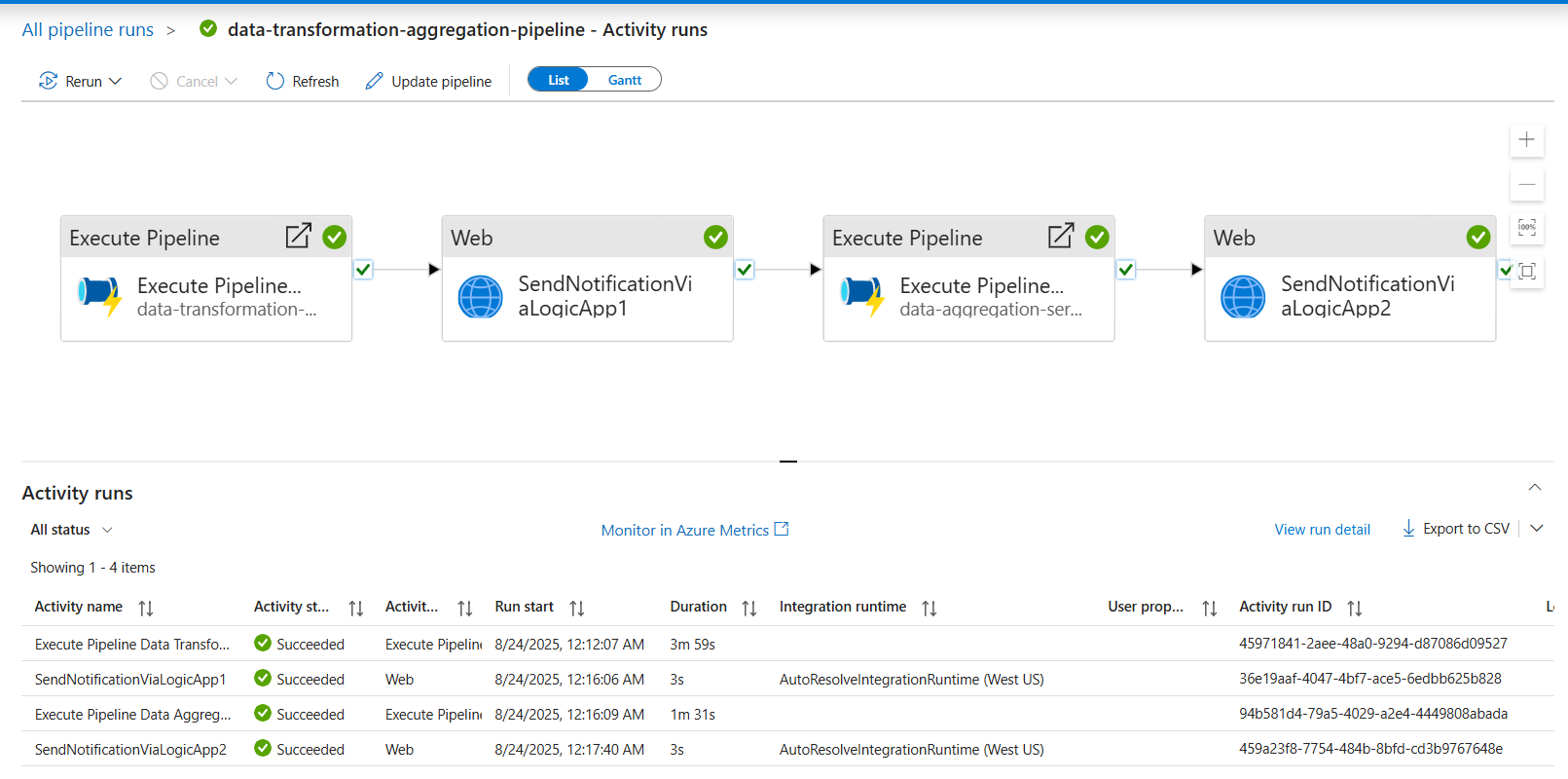




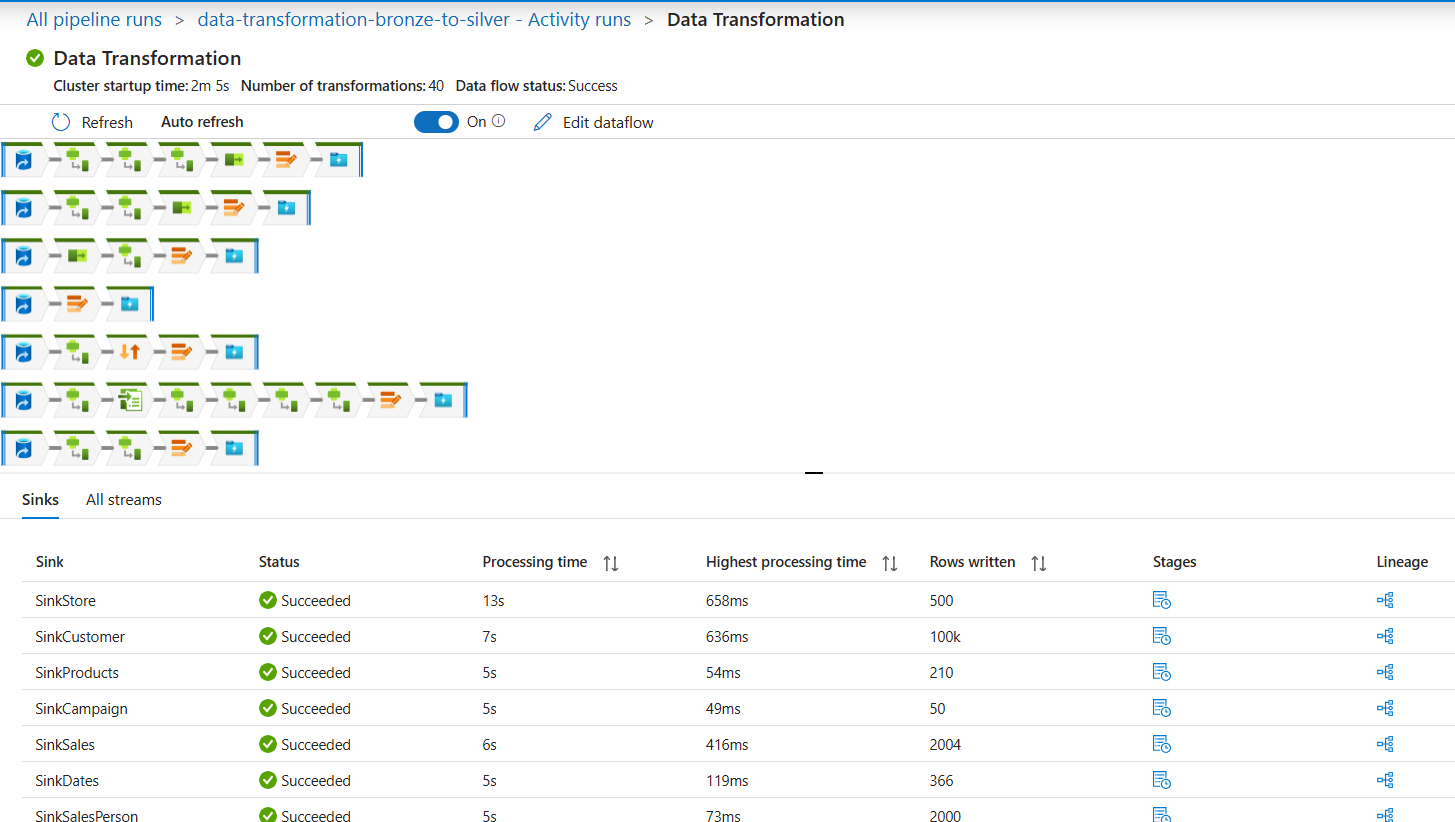


# Data Transformation and Aggregation Pipeline

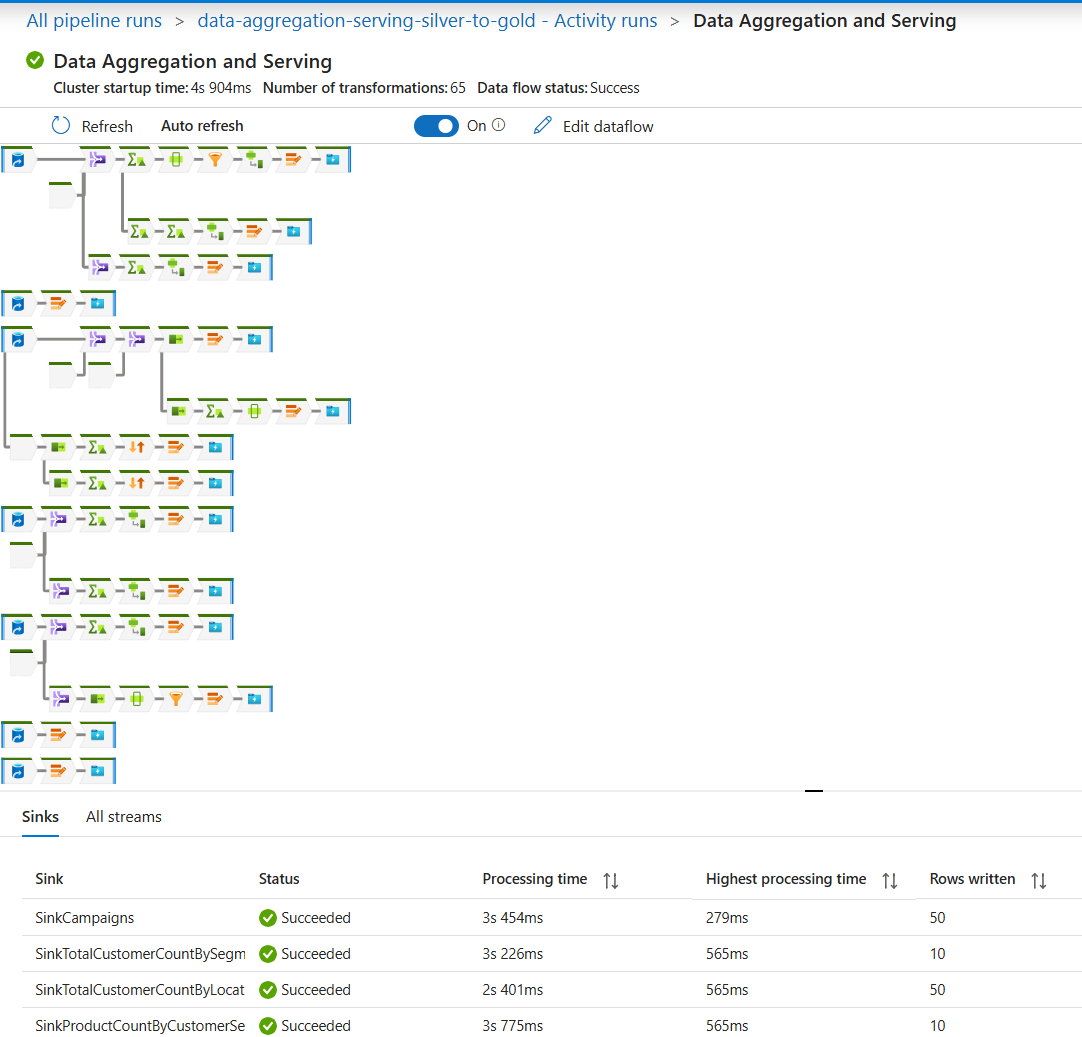
## Parent Pipeline Run



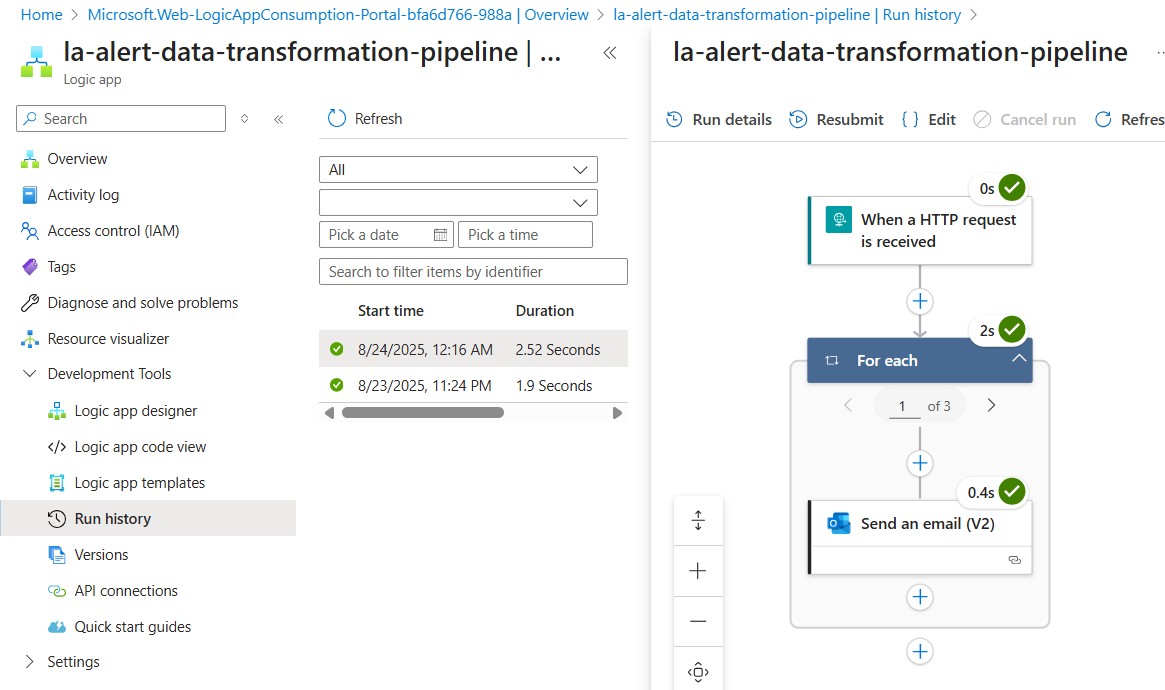
## Data Transformation Pipeline Run

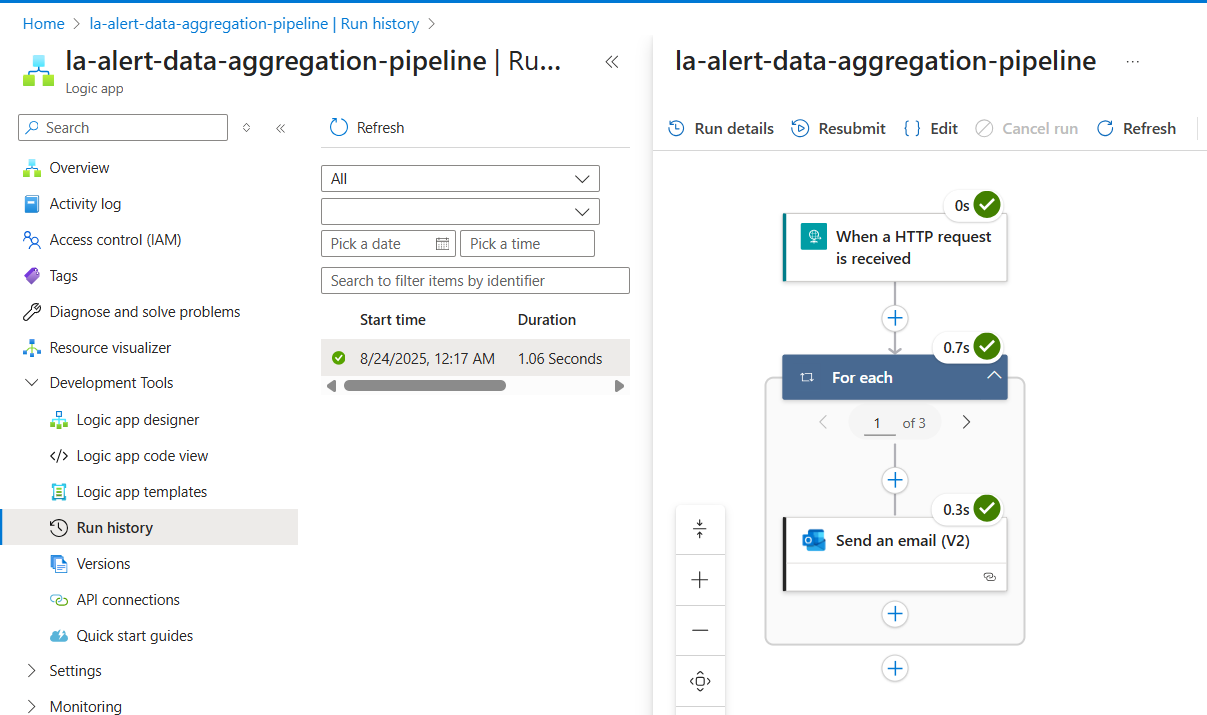


## Data Aggregation Pipeline Run

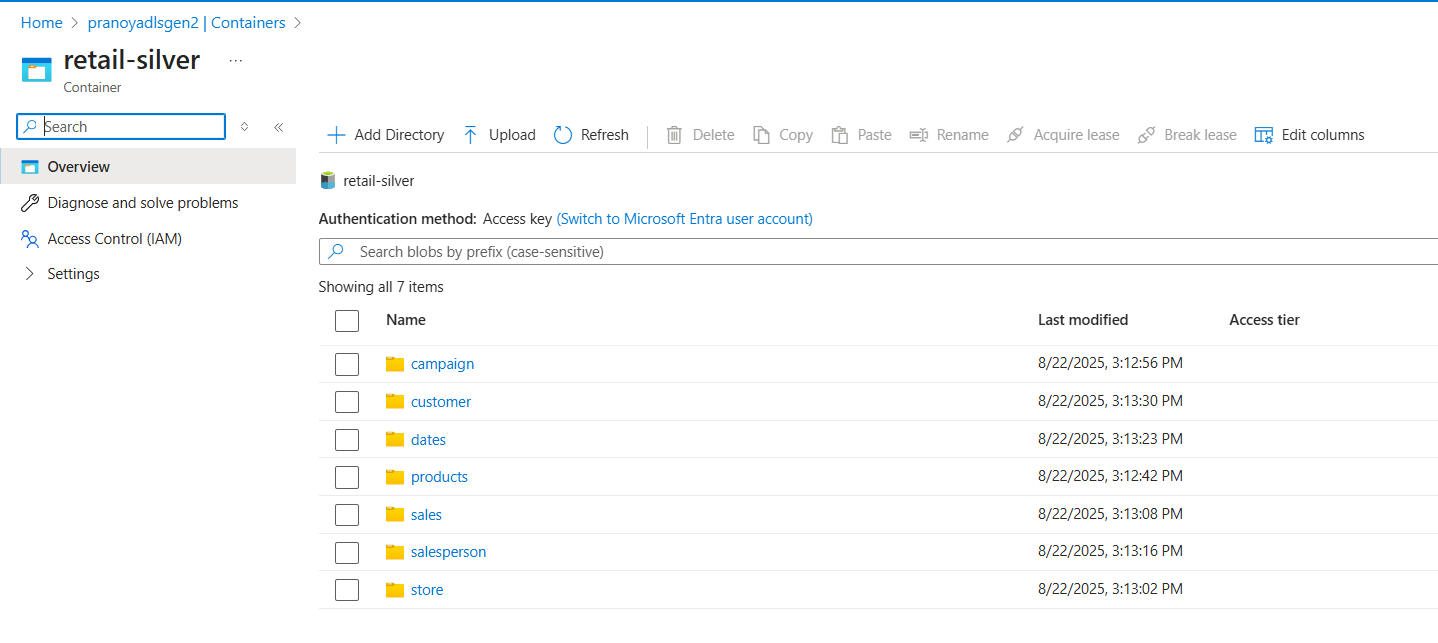


## Alert Notification using Azure Logic App

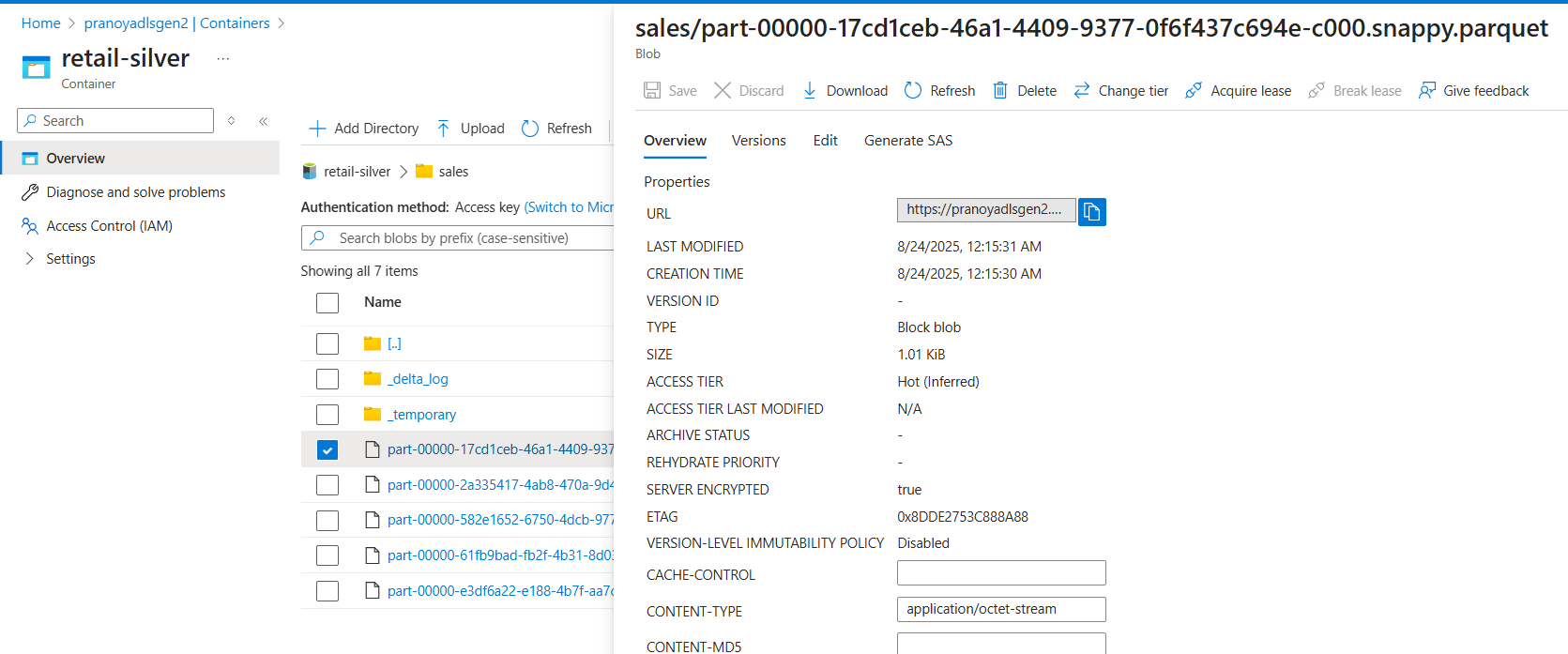




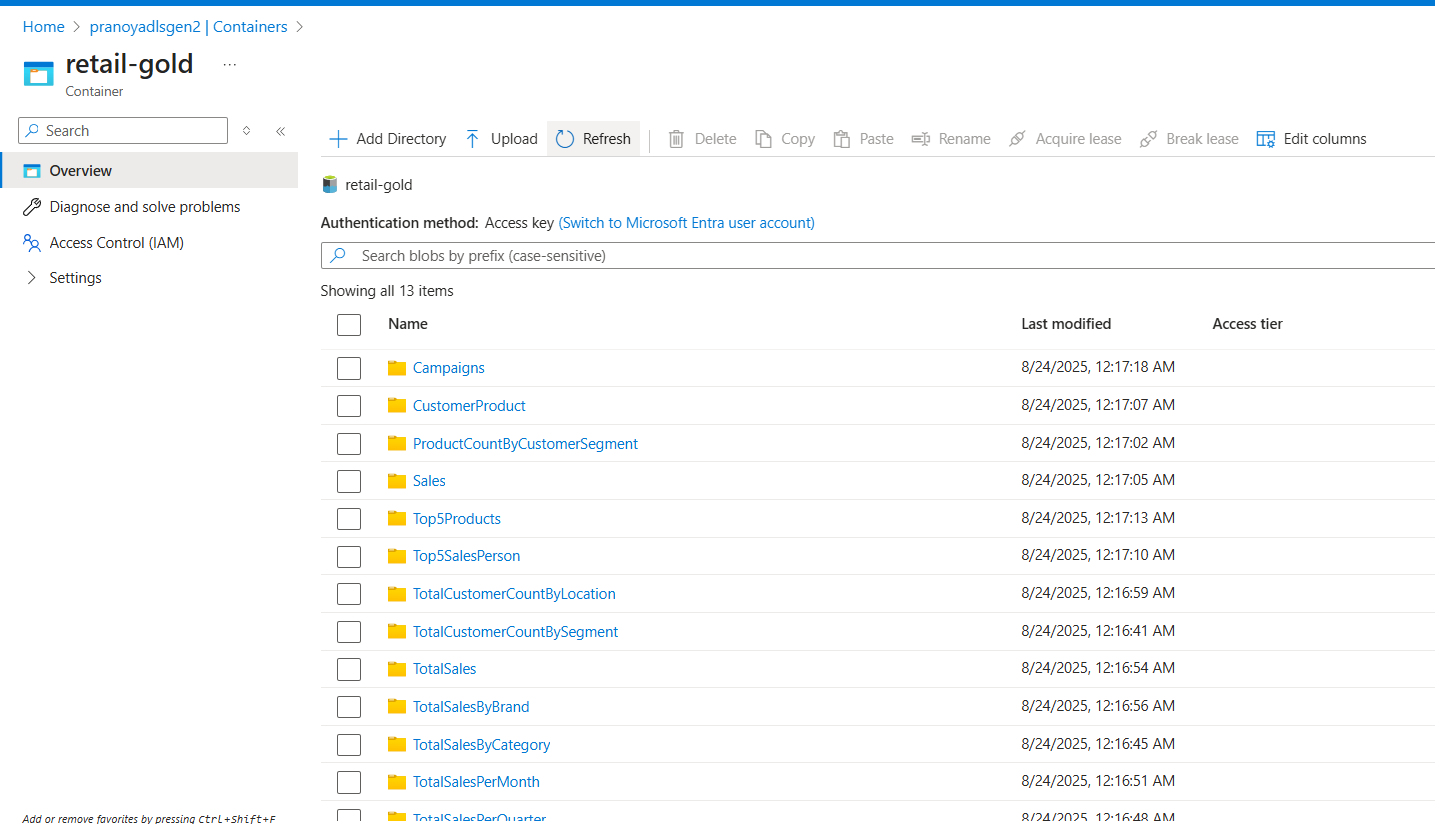
## Cleaned, Transformed Data stored in Silver Layer in Azure Data Lake



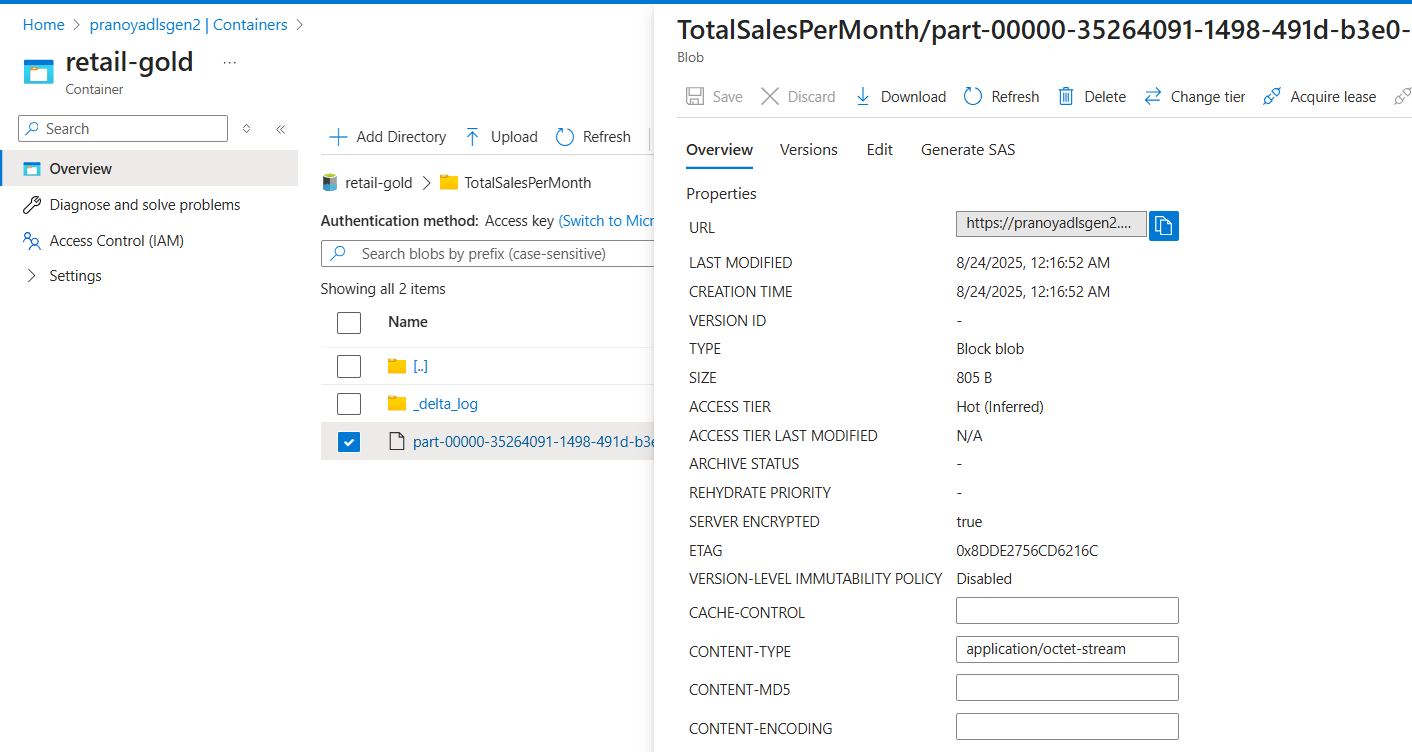
Data stored in Delta Format



## Aggregated, Business-ready Data stored in Gold Layer in Azure Data Lake



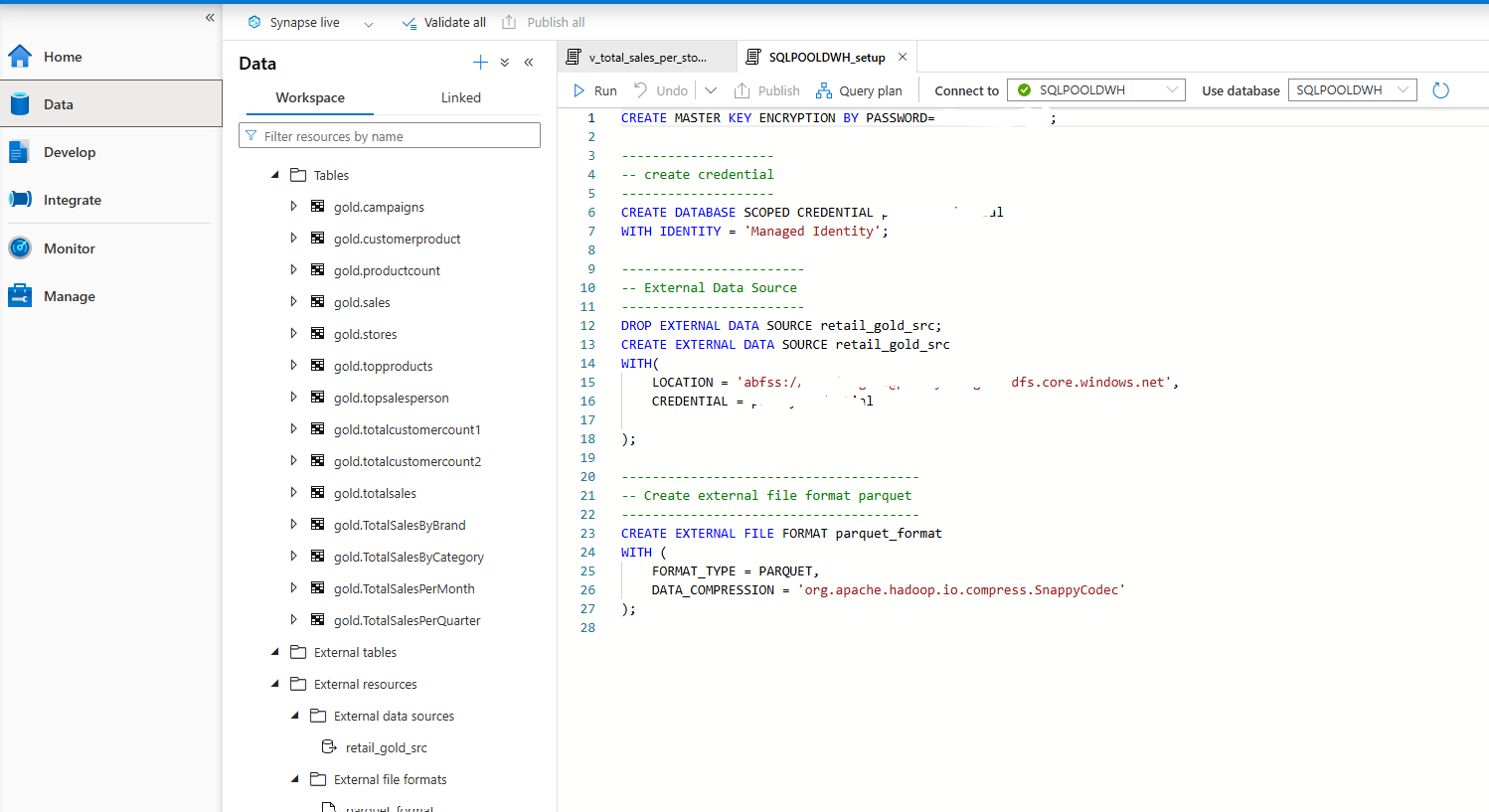
Data stored in Delta Format



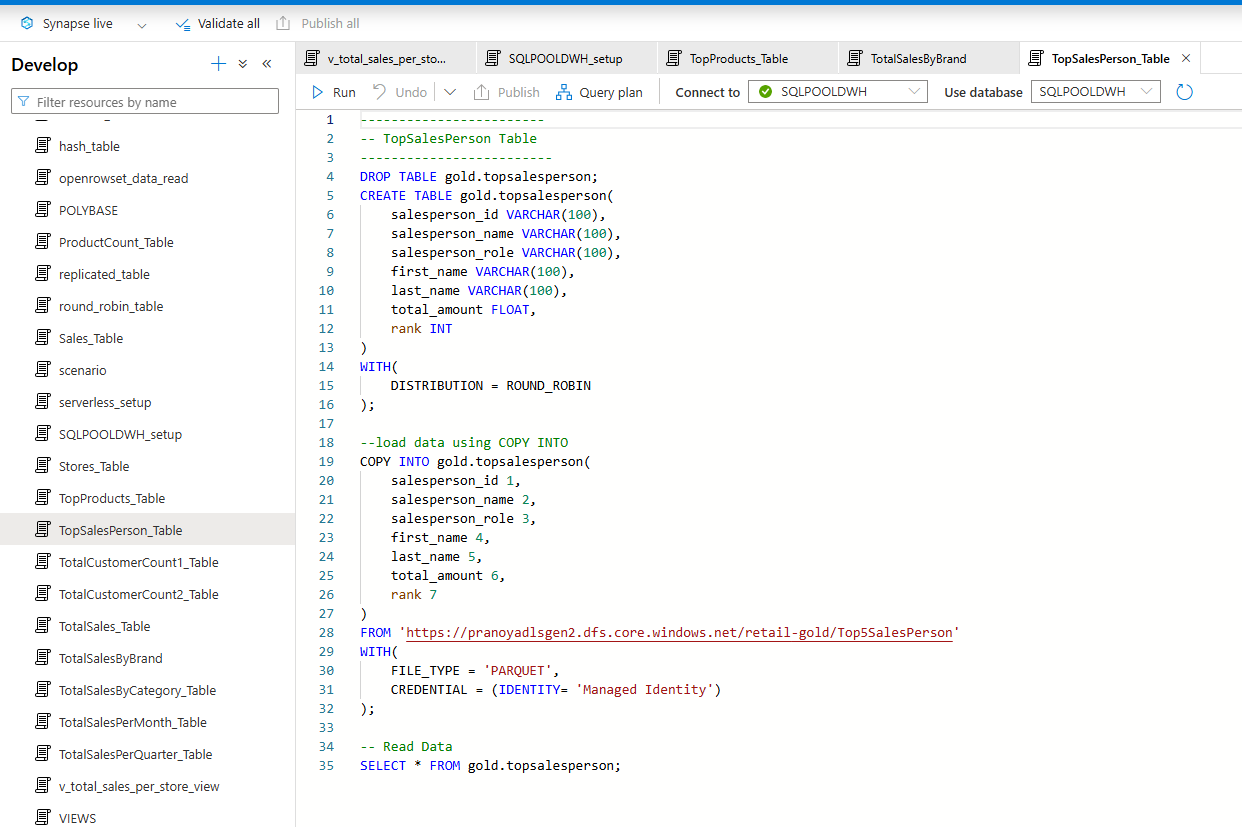
# Data Warehousing and Serving using Azure Synapse Analytics

Setup **Dedicated SQL Pool** and complete the setup

* Create MASTER KEY ENCRIPTION
* Create CREDENTIAL
* Create EXTERNAL DATA SOURCE
* CREATE EXTERNAL FILE FORMAT



Use COPY INTO to copy data from Azure Data Lake Gold Layer to Azure Synapse Analytics Dedicated SQL Pool



Query Data using SQL Server Management Studio

