

DATA ENGINEERING PROJECT

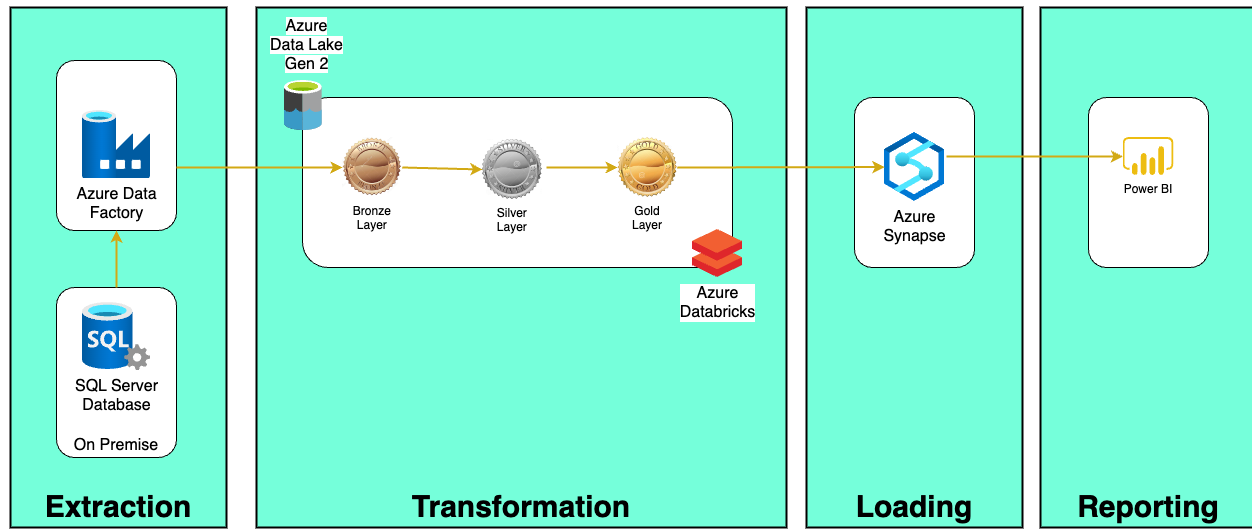
REAL-TIME AZURE END-TO-END DATA PIPELINE PROJECT

By Randy A. Velasco, Data Engineer/Analyst

I. Introduction

Welcome to my Real-Time Data Pipeline Portfolio Project, a testament to my commitment to harnessing the power of data in real-time for informed decision-making. In this project, I showcase a sophisticated end-to-end solution that seamlessly ingests, transforms, and analyzes data from an on-premise SQL Server database, all while utilizing the robust capabilities of Microsoft Azure. This project is designed as a real-time data pipeline with automatic triggers, ensuring agility and responsiveness to evolving business needs.

II. Project Overview



Real-Time Azure End-to-End Data Pipeline Project

1. Real-Time Data Ingestion with Azure Data Factory:

- Configuring automatic triggers for timely data extraction from on-premise SQL Server to Azure Data Lake.
- Ensuring continuous and efficient data flow using Azure Data Factory pipelines.

2. Dynamic Data Transformation with Azure Databricks:

- Leveraging the power of Azure Databricks for real-time, dynamic data transformations.
- Adapting to evolving data requirements and ensuring agility in response to changing business conditions.

3. Immediate Data Loading into Azure Synapse Analytics:

- Utilizing Azure Synapse Analytics for real-time data loading and near-instantaneous availability.
- Harnessing the scalability and performance of Synapse Analytics for seamless integration with the data pipeline.

4. Automated Visualization with Microsoft Power BI:

- Configuring automatic triggers for refreshing Power BI dashboards in real-time.
- Empowering users with up-to-the-minute insights through the seamless integration of Power BI with Azure Synapse Analytics.

5. Governance and Monitoring with Azure Services:

- Leveraging Azure Key Vault for safeguarding sensitive information, ensuring compliance, and enabling automatic authentication.
- Configuring real-time monitoring and alerting mechanisms for proactive governance.

This project exemplifies my dedication to building data pipelines that go beyond traditional batch processing, embracing the dynamics of real-time data. The incorporation of automatic triggers ensures that insights are delivered promptly, enhancing the agility and responsiveness of organizations in an ever-changing business landscape

III. Tools Showcased

This Real-Time Data Pipeline Project incorporates a suite of powerful tools, each playing a pivotal role in achieving a seamless, end-to-end data processing workflow. Here's a glimpse of the tools showcased in this project:

1. MS SQL – On Premise

- Role: Source Database
- Significance: The on-premise MS SQL Server serves as the source of raw data, and our pipeline ensures efficient extraction and utilization of this data.

2. Azure Data Factory

- Role: Data Ingestion and Orchestration
- Significance: Azure Data Factory facilitates the seamless movement of data from on-premise SQL Server to Azure Data Lake. Its automatic triggers ensure timely and reliable data ingestion.

3. Azure Synapse

- Role: Data Warehousing and Processing
- Significance: Azure Synapse Analytics provides a robust platform for data storage, processing, and analysis. It seamlessly integrates with our pipeline for efficient data loading and near-instantaneous availability.

4. Azure Databricks

- Role: Real-Time Data Transformation
- Significance: Azure Databricks empowers our pipeline with dynamic, real-time data transformations. It ensures agility in adapting to evolving data requirements, enhancing the overall efficiency of the data processing workflow.

5. Azure Key Vault

- Role: Security and Compliance
- Significance: Azure Key Vault is employed to securely store sensitive information, such as authentication credentials and encryption keys. It ensures compliance and enhances security in the data processing pipeline.

6. Azure Data Lake

- Role: Data Storage
- Significance: Azure Data Lake serves as a scalable and secure storage repository for raw and processed data. It plays a crucial role in the efficient management of data throughout the pipeline.

7. MS Power BI

- Role: Data Visualization and Analytics
- Significance: Microsoft Power BI is utilized to create interactive dashboards, providing users with real-time insights. Its integration with Azure Synapse Analytics ensures immediate access to the latest data.

This curated selection of tools demonstrates my commitment to leveraging the best-in-class technologies for building robust, real-time data pipelines. Together, they form a cohesive ecosystem that empowers organizations to unlock the full potential of their data for strategic decision-making and business success.

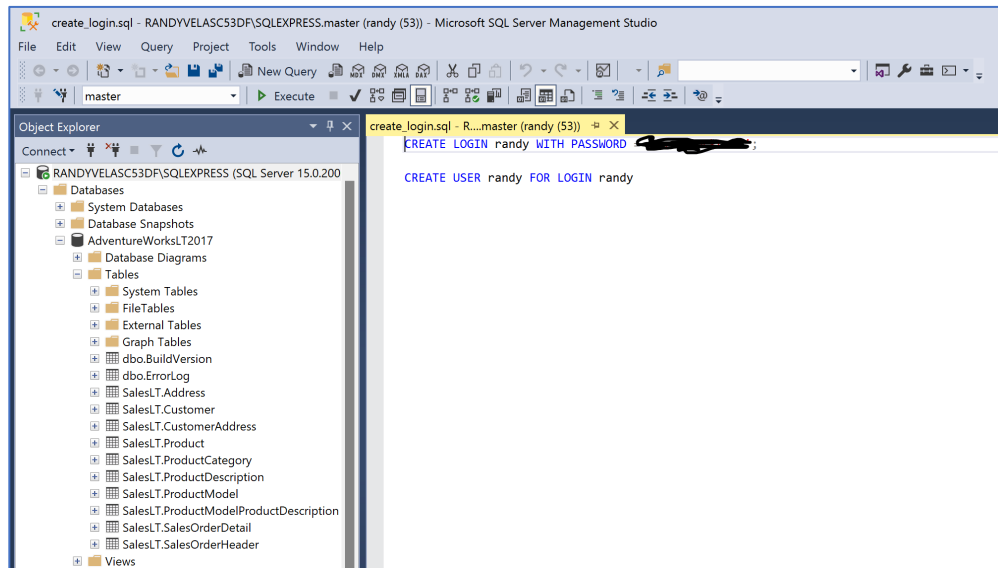
IV. Procedures Performed in Real-Time Data Pipeline Project

This Real-Time Data Pipeline Project involved a meticulously planned series of procedures to ensure a seamless and efficient end-to-end data processing workflow. Here's an overview of the key procedures performed:

1. Environmental Setup (MS SQL and Azure)

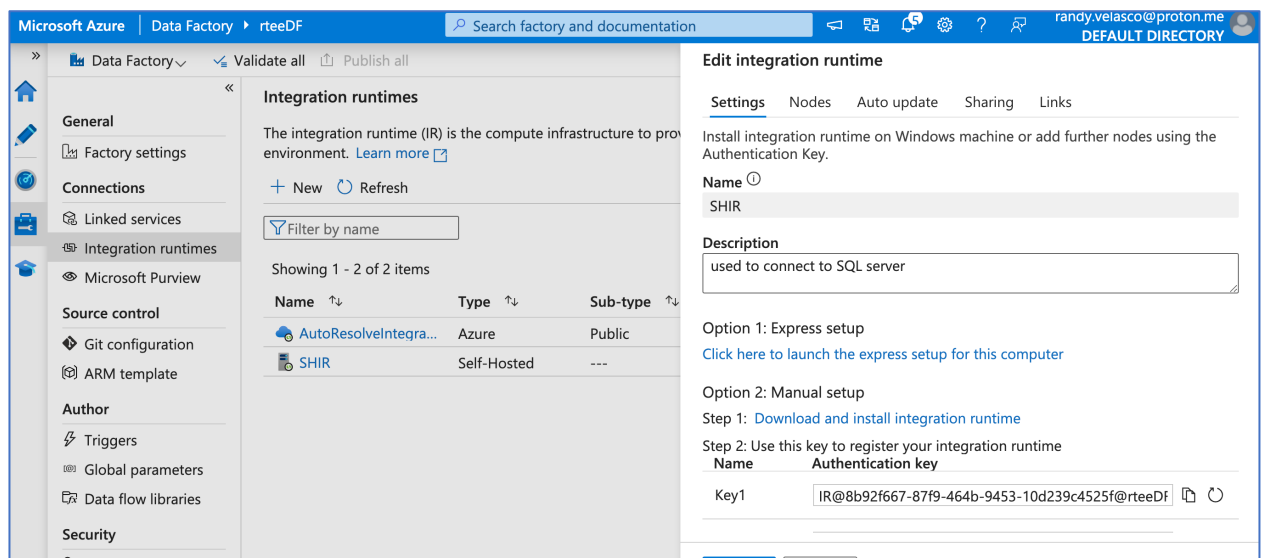
a. Configure Access to On-Premise SQL Server:

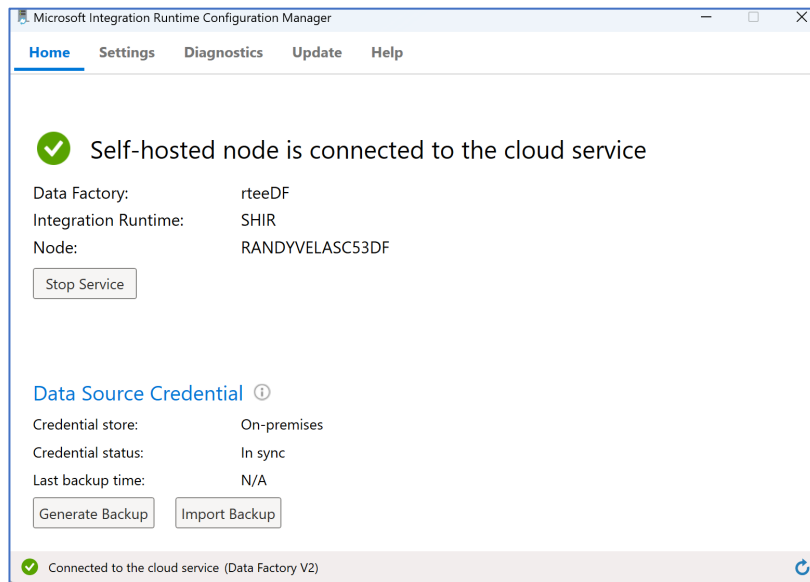
- Facilitated secure and authorized access to the on-premise MS SQL Server, ensuring a controlled connection for data extraction.



b. Setup SQL Integration Runtime:

- Configured SQL Integration Runtime to enable the smooth transfer of data between on-premise SQL Server and Azure services.

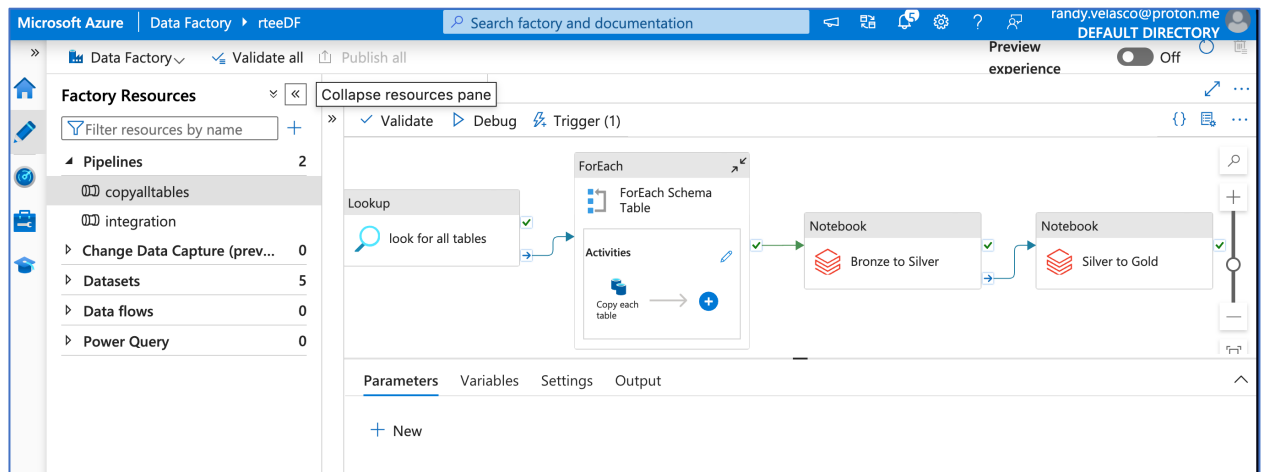




2. Data Ingestion

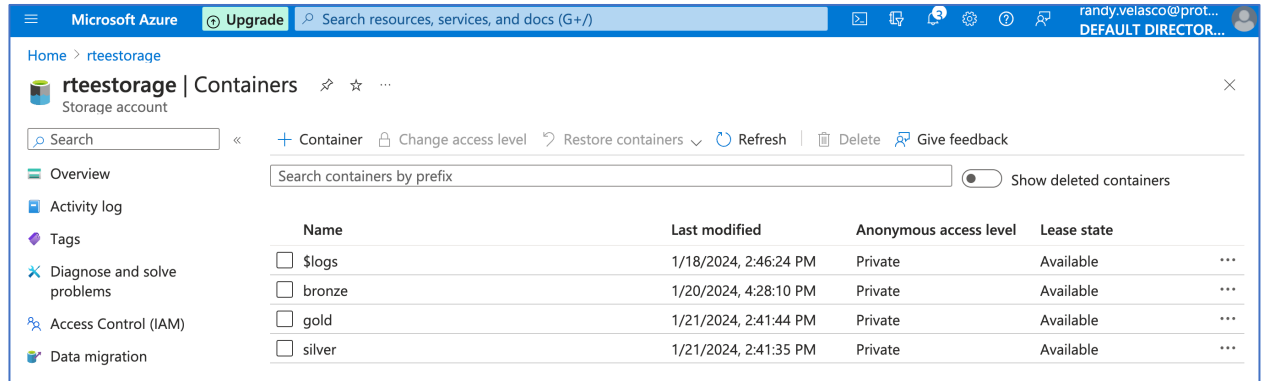
a. Create Automated Pipeline for SQL Server to Azure Data Lake:

- Established an automated pipeline using Azure Data Factory to extract tables from the on-premise SQL Server and load them into Azure Data Lake.

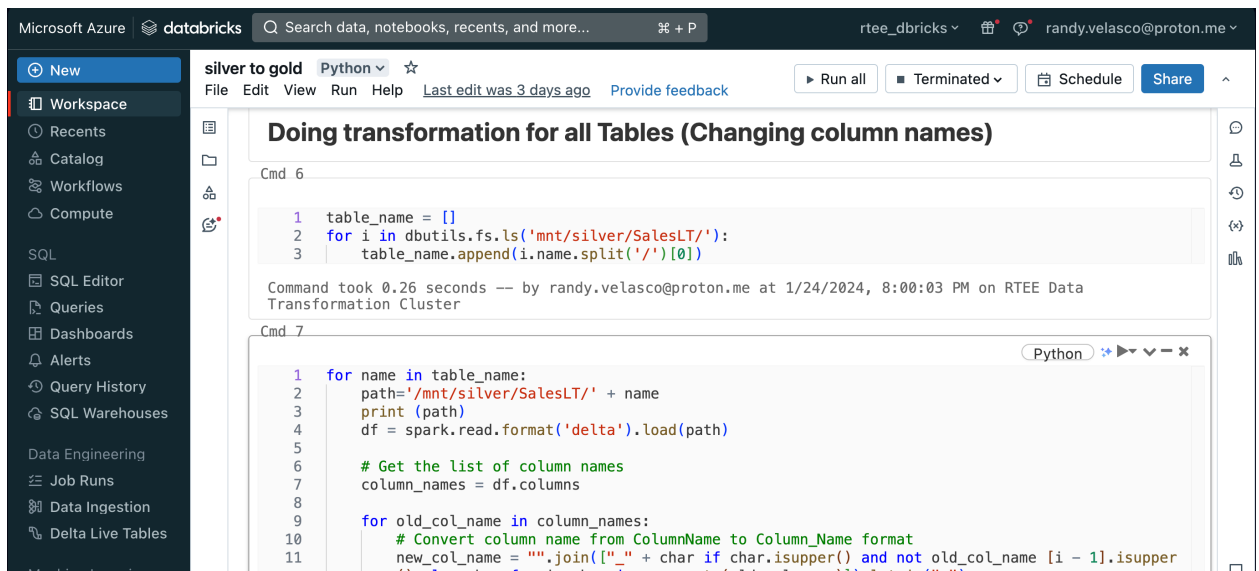


3. Data Transformation

- a. Create Separate Containers for Raw, Semi-Transformed, and Final Transformed Data:
 - Implemented a structured storage system with distinct containers for raw data (bronze), semi-transformed data (silver), and final transformed data (gold).



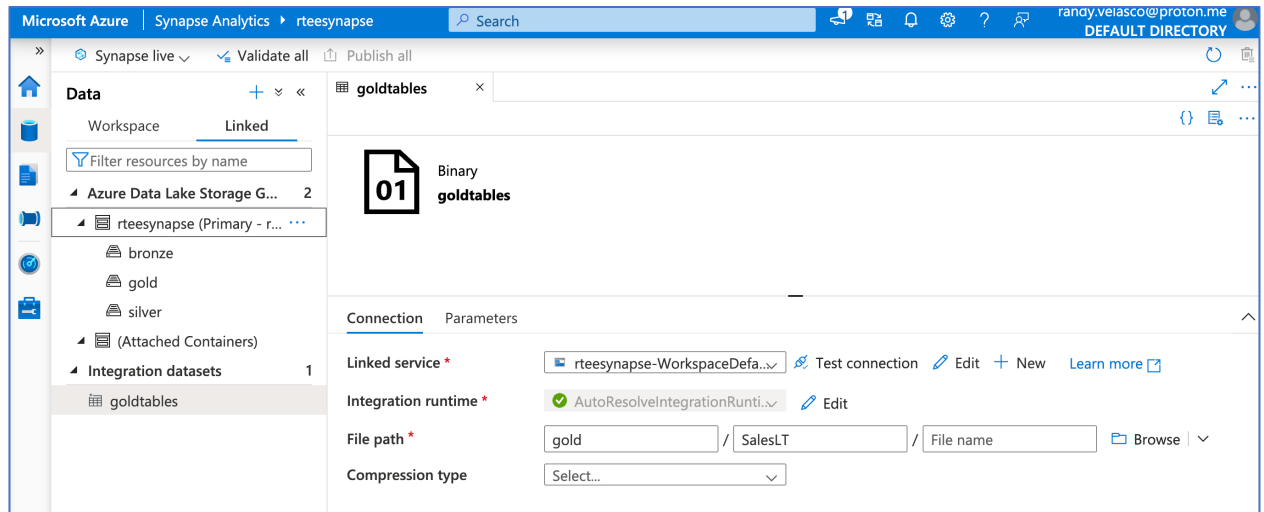
- b. Perform Data Cleaning and Transformation using PySpark:
 - Leveraged PySpark within Azure Databricks to conduct data cleaning and transformation activities, ensuring the data meets quality and format standards.



4. Data Loading

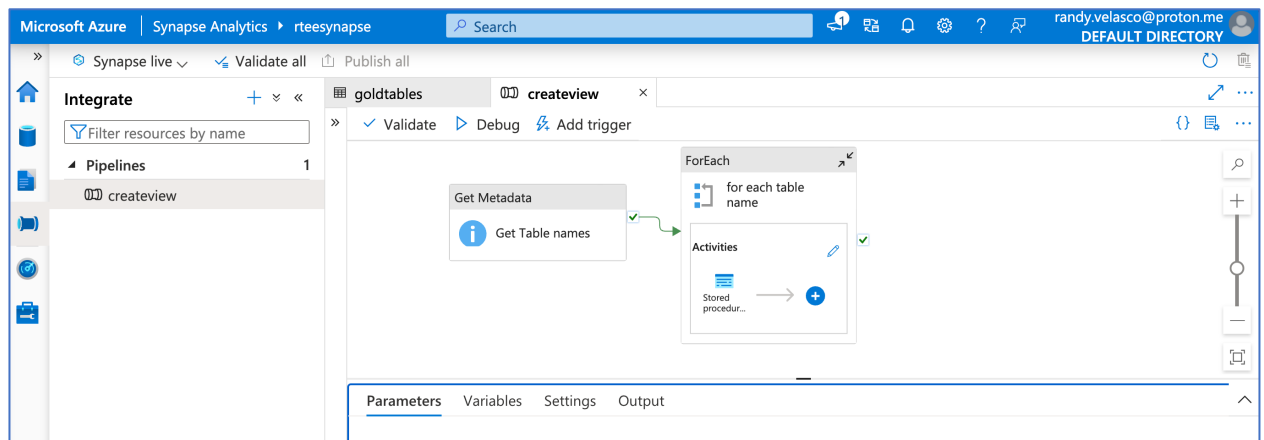
a. Link Azure Synapse with Transformed Data in Data Lake:

- Established a connection between Azure Synapse Analytics and the transformed data stored in Azure Data Lake for seamless data loading.



b. Create Views from Tables:

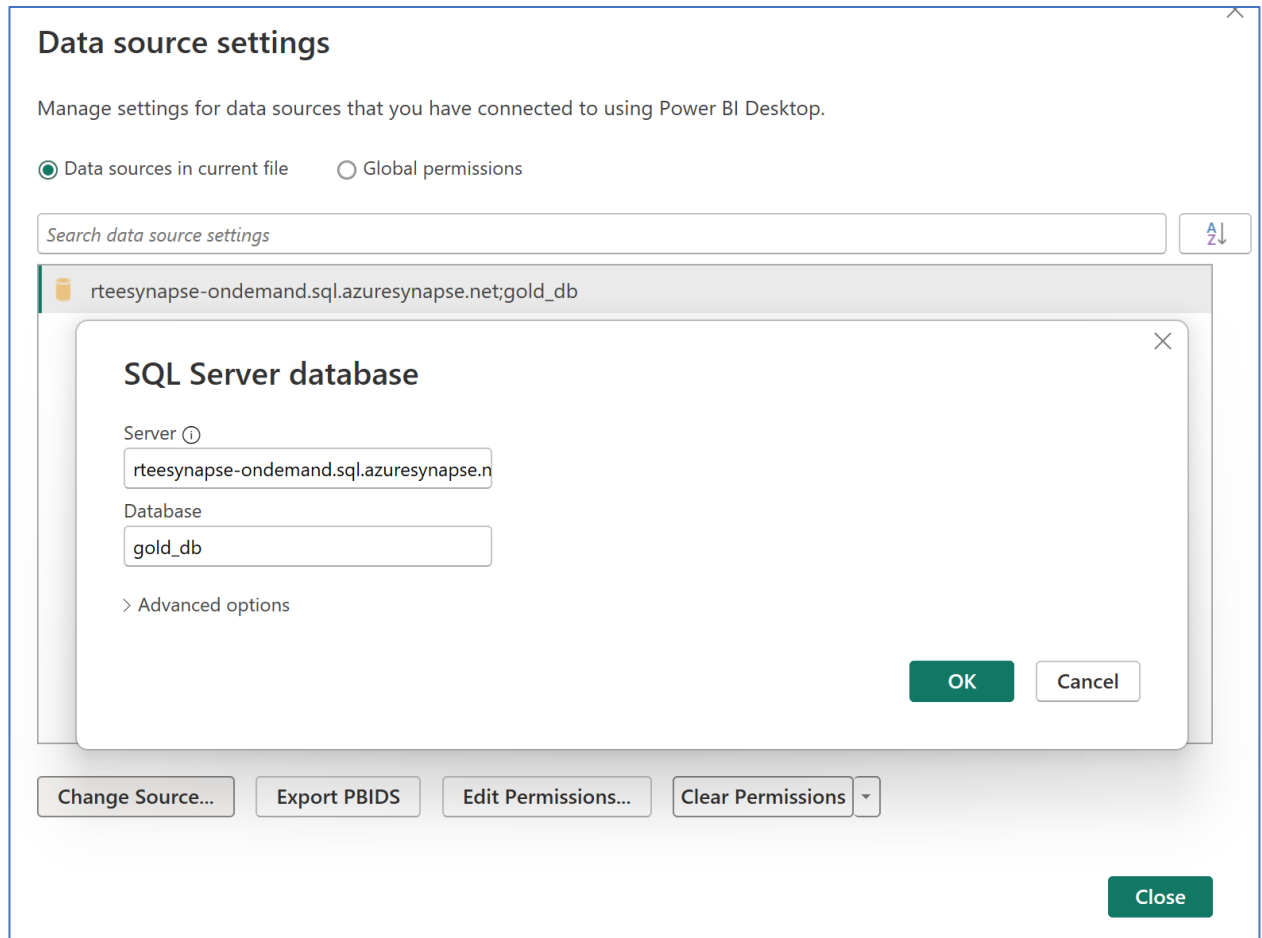
- Utilized Azure Synapse Analytics to create views from the transformed data, providing a structured and efficient way to query the data.



5. Reporting

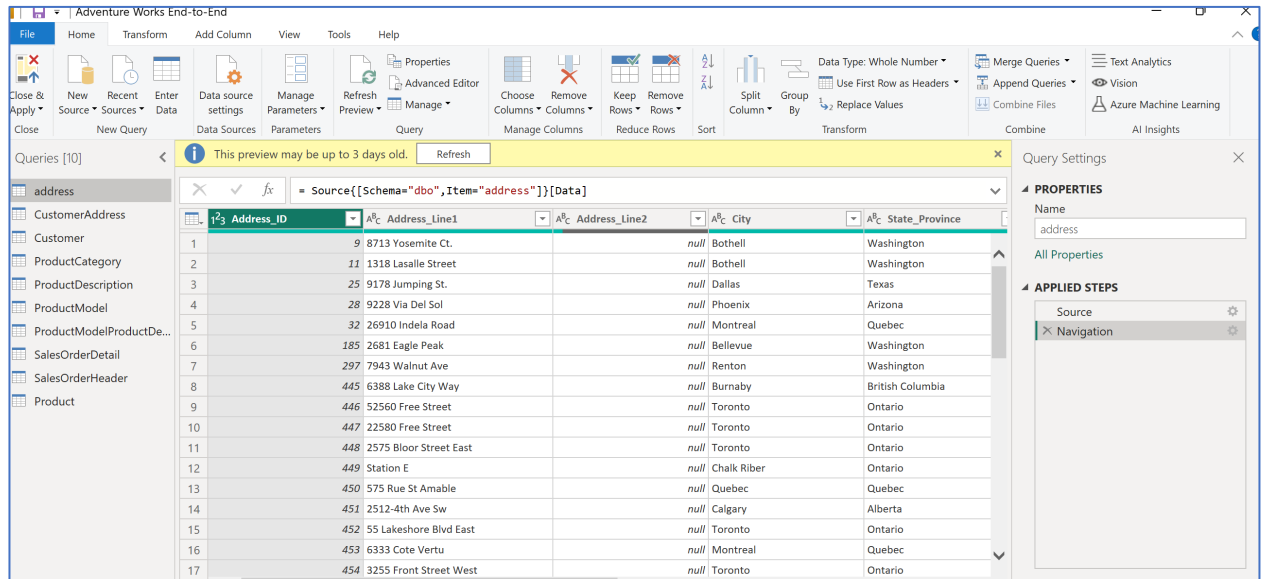
a. Establish Data Connection from Power BI to Azure Synapse:

- Configured a secure data connection between Microsoft Power BI and Azure Synapse Analytics, ensuring real-time access to the latest data.



b. Extract Tables from Azure Synapse to Power BI:

- Extracted relevant tables and data sets from Azure Synapse Analytics, making them available for visualization and analysis in Power BI.



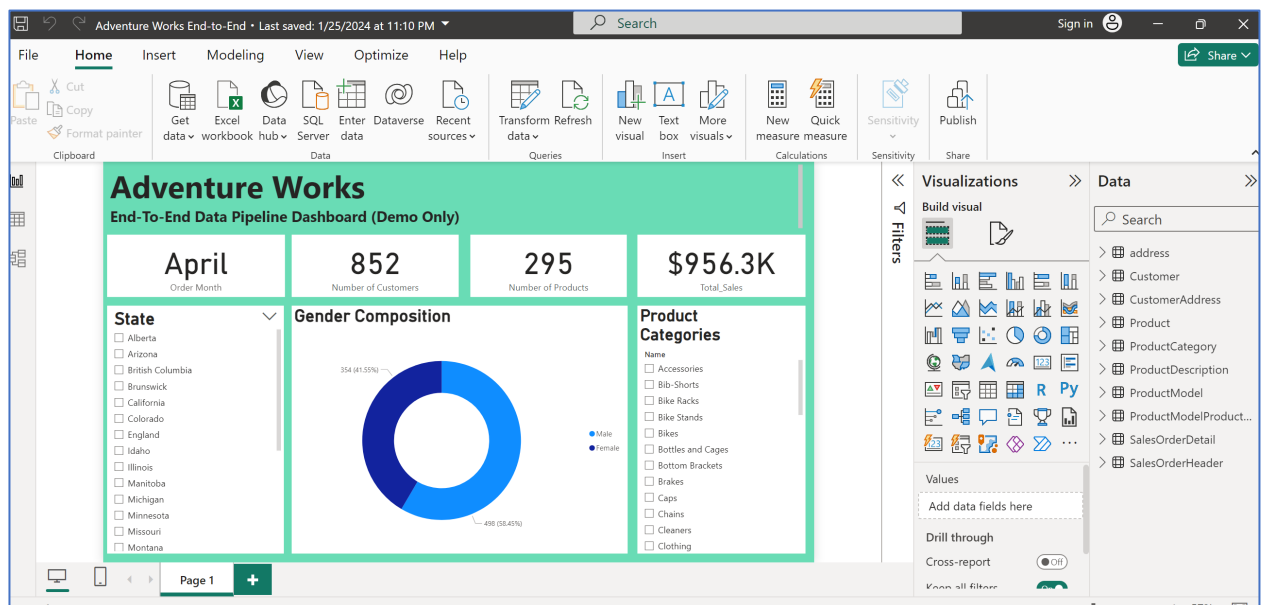
Query: address

Source: Source([Schema="dbo",Item="address"])[Data]

Address_ID	Address_Line1	Address_Line2	City	State_Province
1	8713 Yosemite Ct.		Bothell	Washington
2	1318 Lasalle Street		Bothell	Washington
3	9178 Jumping St.		Dallas	Texas
4	9228 Via Del Sol		Phoenix	Arizona
5	26910 Indella Road		Montreal	Quebec
6	2681 Eagle Peak		Bellevue	Washington
7	7943 Walnut Ave		Renton	Washington
8	6388 Lake City Way		Burnaby	British Columbia
9	52560 Free Street		Toronto	Ontario
10	22580 Free Street		Toronto	Ontario
11	2575 Bloor Street East		Toronto	Ontario
12	Station E		Chalk River	Ontario
13	575 Rue St Amable		Quebec	Quebec
14	2512-4th Ave Sw		Calgary	Alberta
15	55 Lakeshore Blvd East		Toronto	Ontario
16	6333 Cote Vertu		Montreal	Quebec
17	3255 Front Street West		Toronto	Ontario

c. Create Dashboard:

- Developed a comprehensive dashboard in Microsoft Power BI, providing stakeholders with a visually intuitive interface for data exploration and decision-making.



6. Scheduling Automatic Triggers

a. Implement Automatic Trigger for Pipeline Execution:

Configured scheduled automatic triggers to initiate the data pipeline at predefined intervals. This automation ensures the seamless execution of the entire data processing workflow without manual intervention.

Edit trigger

Name *

Description

Type *

ScheduleTrigger

Start date * ⓘ

Time zone * ⓘ

Kuala Lumpur, Singapore (UTC+8) ▼

Recurrence * ⓘ

Every

Day(s) ▼

▼ Advanced recurrence options

Execute at these times ⓘ

Hours

 ×

Minutes

 ×

Schedule execution times

22:12

☐ Specify an end date

Annotations

+ New

Status ⓘ

☒ Started ☐ Stopped

b. Logging and Monitoring of Trigger Events:

Implemented logging mechanisms to track and monitor the execution of automatic triggers, enabling proactive identification and resolution of any potential issues.

Trigger name	Trigger type	Trigger time	Status	Pipelines	Run	Message
scheduled_trigger	Schedule trigger	1/25/2024, 10:11:5	Succeeded	1	Original	

These procedures collectively demonstrate the meticulous orchestration of tasks involved in setting up, ingesting, transforming, loading, and reporting within the real-time data pipeline. The result is a robust and agile system capable of delivering timely insights for informed decision-making.

(END OF DOCUMENT)