Microsoft Azure Data Engineering Training Course

Capstone Project

Data-Driven Decision-Making for Enhanced Online Retail Experience and Increased Sales

Objectives:-

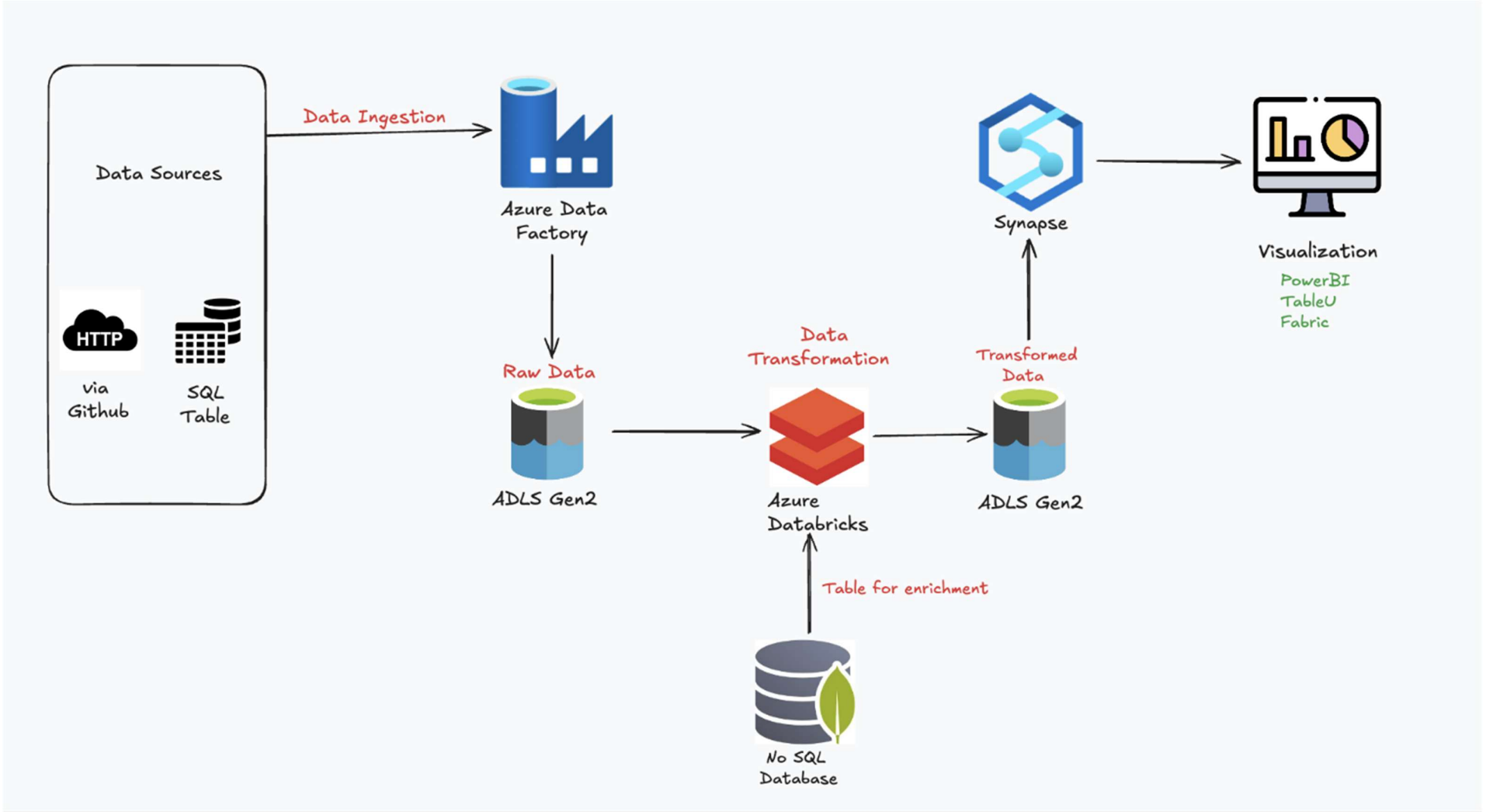
1. **Implement Batch Processing in Azure Data Engineering Services for Online Retails Dataset.**
2. **Implement Stream Processing in Azure Data Engineering Services for Online Retails Streaming Dataset.**

Solution for Implementing

Batch Processing

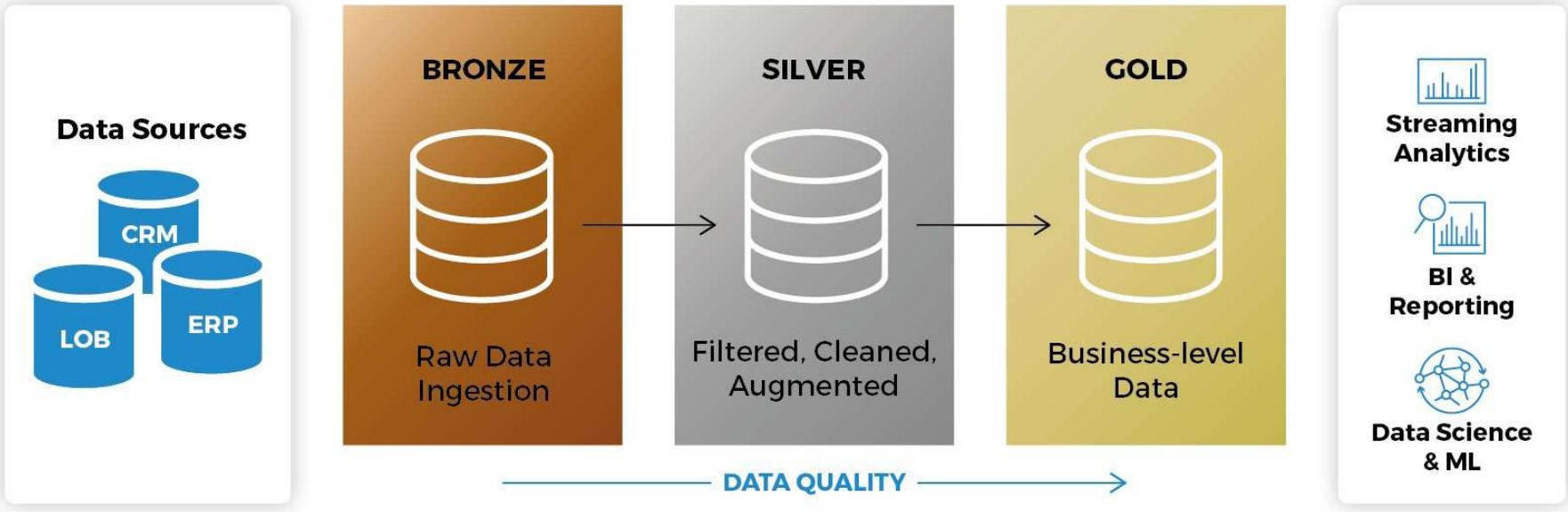
in Azure Data Engineering

Overall Batch processing architecture:-



Medallion Architecture :-

- The Medallion architecture structures data in a multi-tier approach —bronze, silver and gold tier— taking into account and encouraging data quality as it moves through the transformation process (from raw data to valuable business insights).



- **Bronze layer**: This phase marks the input of raw data, which is stored as it is collected, usually from a variety of sources and in formats such as CSV or JSON. The data is usually raw data and varies in quality and structure.

- **Silver layer**: At this point, the data is processed and transformed to achieve cleaner, more structured data. Tasks such as filtering, validation and normalisation of the data are carried out and stored in efficient formats. This phase may include defined schemas and additional metadata.

- **Gold layer**: This stage contains data already prepared for analysis and business use. In the Gold layer, advanced transformations and aggregations are performed to create rich data sets. The data is structured, optimised for fast queries and can be enriched with additional information or merged with other data sources for deeper insights.

Dataset used in this project:-

Brazilian E-Commerce Public Dataset by Olist:-Welcome! This is a Brazilian ecommerce public dataset of orders made at Olist Store. The dataset has information of 100k orders from 2016 to 2018 made at multiple marketplaces in Brazil. Its features allows viewing an order from multiple dimensions: from order status, price, payment and freight performance to customer location, product attributes and finally reviews written by customers. We also released a geolocation dataset that relates Brazilian zip codes to lat/lng coordinates.

Datasets and its details:-

1. Customers Dataset

This dataset has information about the customer and its location. Use it to identify unique customers in the orders dataset and to find the orders delivery location.At our system each order is assigned to a unique customer_id. This means that the same customer will get different ids for different orders. The purpose of having a customer_unique_id on the dataset is to allow you to identify customers that made repurchases at the store. Otherwise you would find that each order had a different customer associated with.

2. Geolocation Dataset

This dataset has information Brazilian zip codes and its lat/lng coordinates. Use it to plot maps and find distances between sellers and customers.

3. Order Items Dataset

This dataset includes data about the items purchased within each order.

Example:

The order_id = 00143d0f86d6fbd9f9b38ab440ac16f5 has 3 items (same product). Each item has the freight calculated accordingly to its measures and weight. To get the total freight value for each order you just have to sum.

The total order_item value is: 21.33 * 3 = 63.99

The total freight value is: 15.10 * 3 = 45.30

The total order value (product + freight) is: 45.30 + 63.99 = 109.29

4. Payments Dataset

This dataset includes data about the orders payment options.

5. Order Reviews Dataset

This dataset includes data about the reviews made by the customers.

After a customer purchases the product from Olist Store a seller gets notified to fulfill that order. Once the customer receives the product, or the estimated delivery date is due, the customer gets a satisfaction survey by email where he can give a note for the purchase experience and write down some comments.

6. Order Dataset

This is the core dataset. From each order you might find all other information.

7. Products Dataset

This dataset includes data about the products sold by Olist.

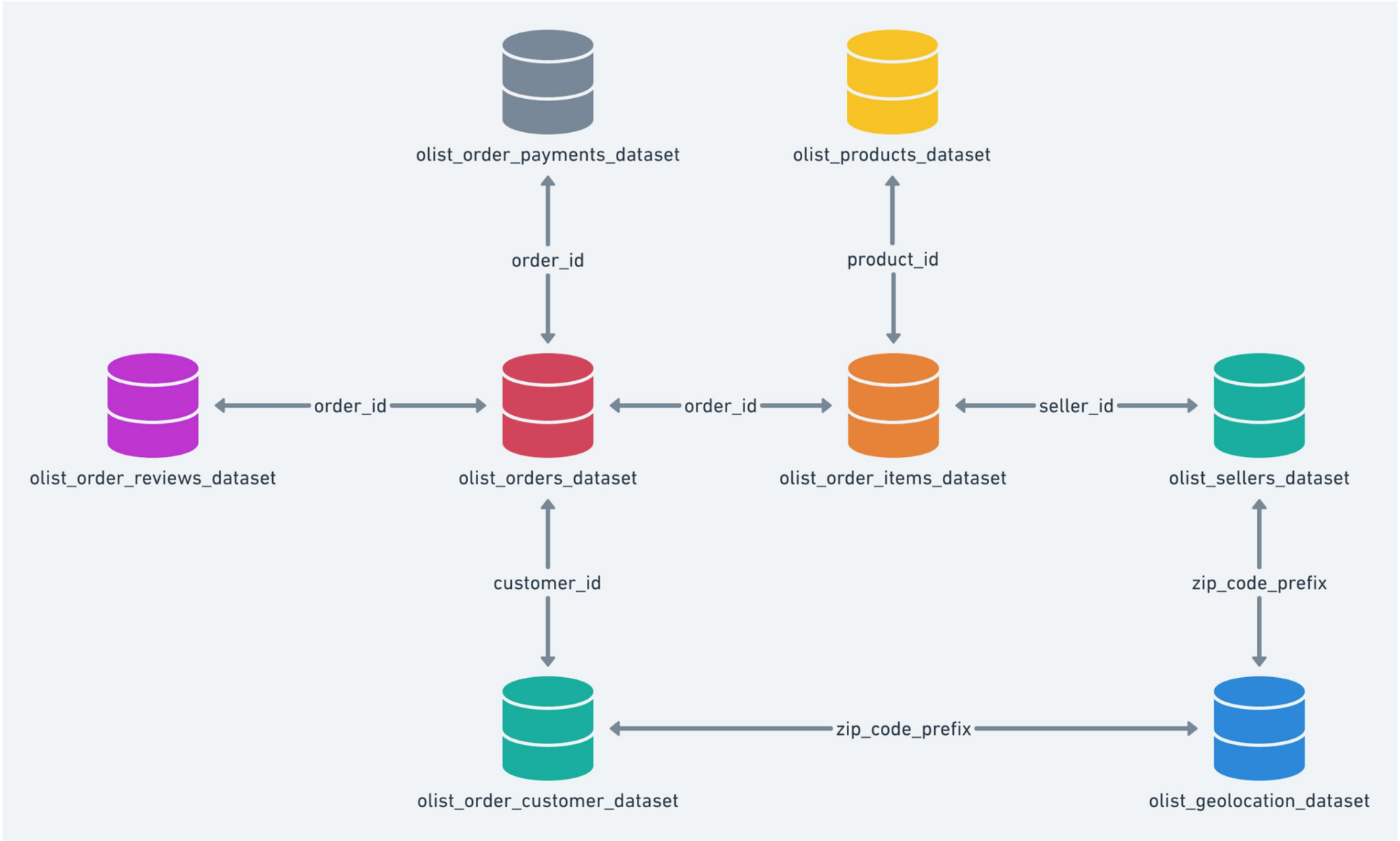
8. Sellers Dataset

This dataset includes data about the sellers that fulfilled orders made at Olist. Use it to find the seller location and to identify which seller fulfilled each product.

9. Category Name Translation

Translates the product_category_name to english.

Data Schema:-


olist_order_payments_dataset
olist_products_dataset
order_id
product_id
order_id
order_id
order_id
seller_id
olist_order_reviews_dataset
olist_orders_dataset
olist_order_items_dataset
olist_sellers_dataset
customer_id
zip_code_prefix
zip_code_prefix
olist_order_customer_dataset
olist_geolocation_dataset

Note:- to get realtime project experience we only first store the data in difference places like Github, MySQL, MongoDB

1) **Datasets stored in Github:-**
   olist_customers_dataset.csv
   olist_geolocation_dataset.csv
   olist_order_items_dataset.csv
   olist_order_payments_dataset.csv
   olist_order_reviews_dataset.csv
   olist_products_dataset.csv
   olist_sellers_dataset.csv

2) **Datasets stored in MySQL:-**
   olist_orders_dataset.csv – this is the orders dataset which I stored in MySQL tables.

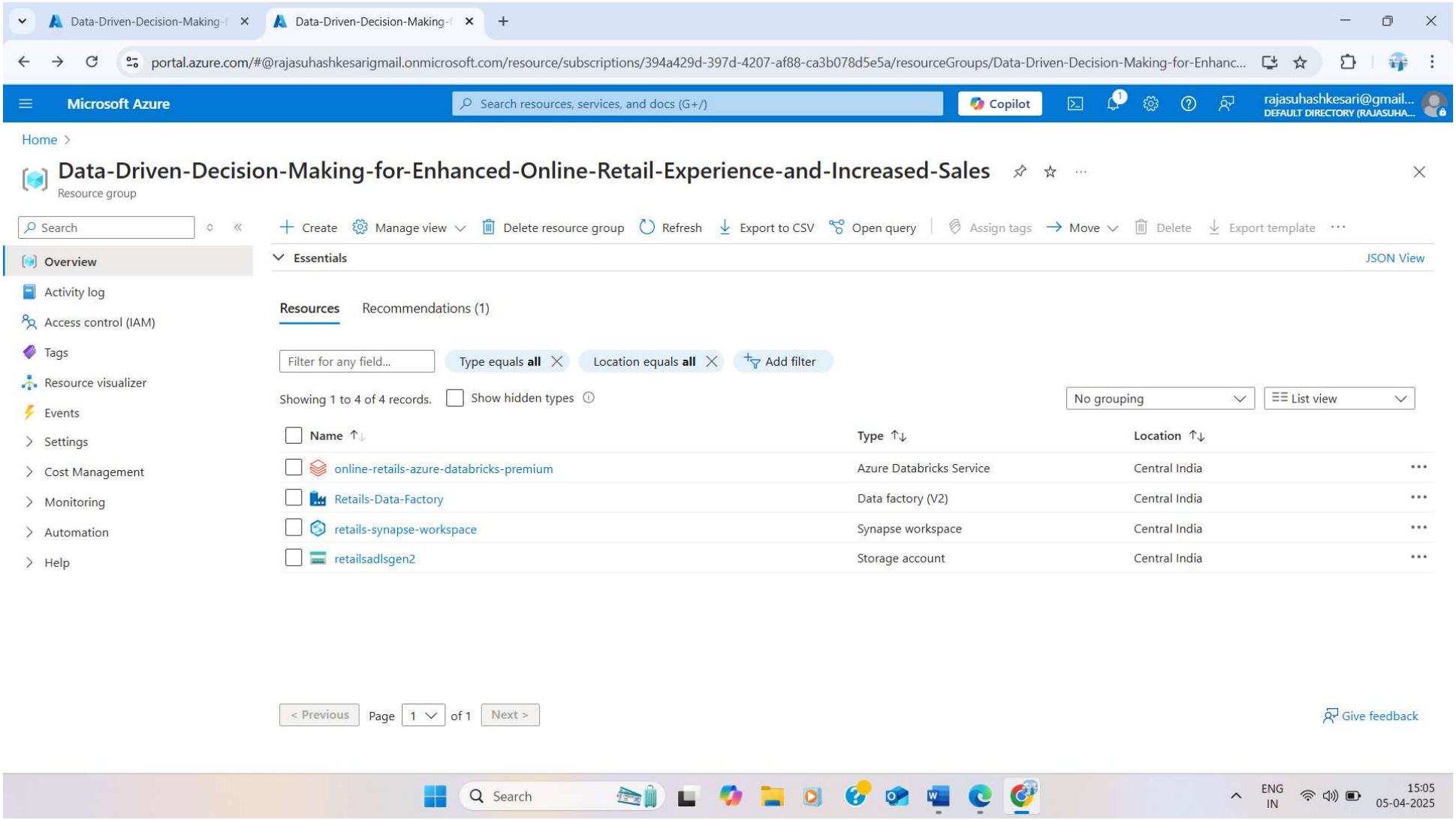
3) **Datasets stored in MongoDB:-**
   product_category_name_translation- this the dataset contains list product Category names and their product category names in English.

Step by Step implementation

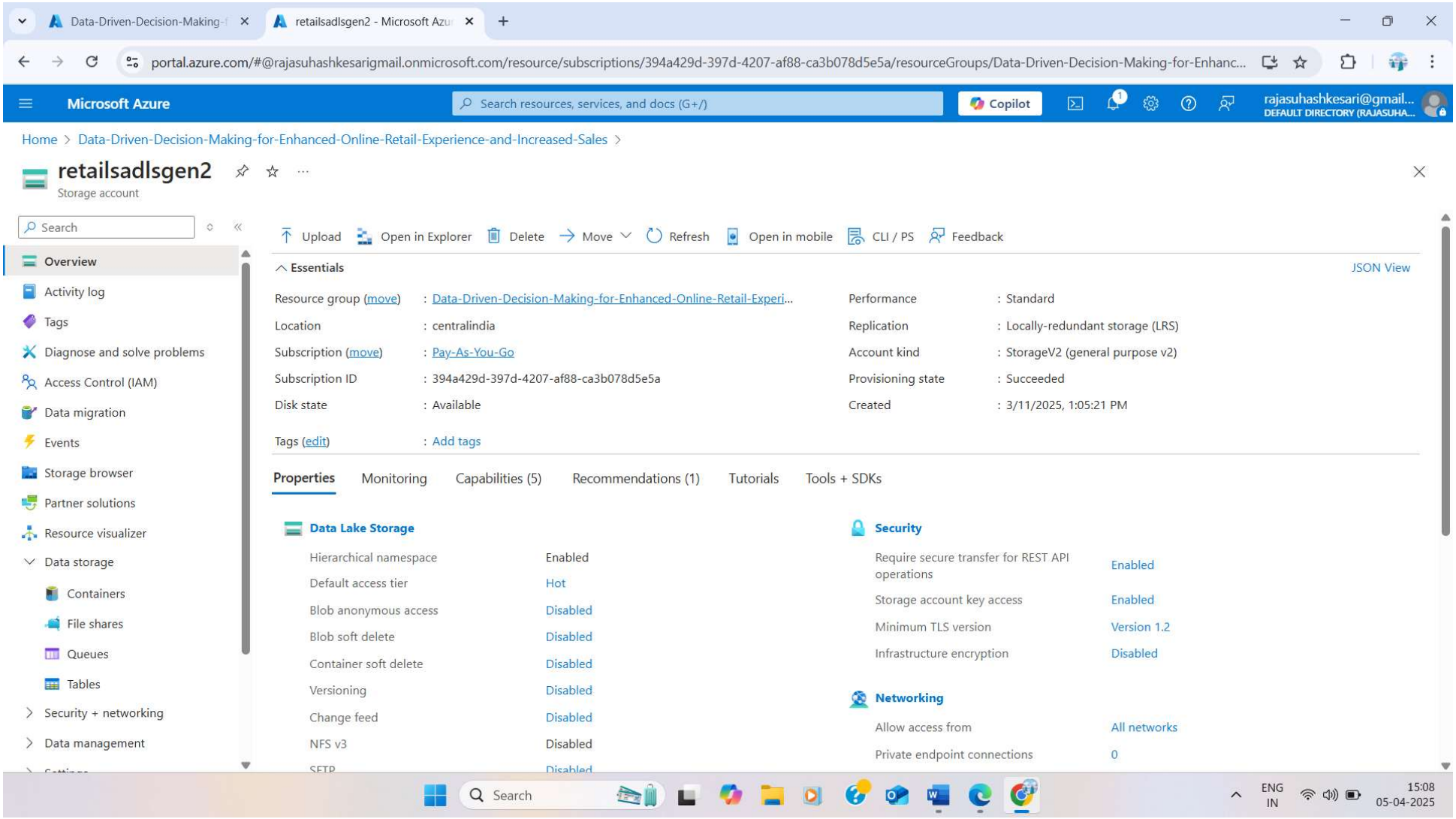
Azure Services Required for Data Engineering:-

1. **Resource Group**
2. **Azure Data Lake Storage Gen2.**
3. **Azure Data Factory.**
4. **Azure Databricks.**
5. **Azure Synapse analytics.**

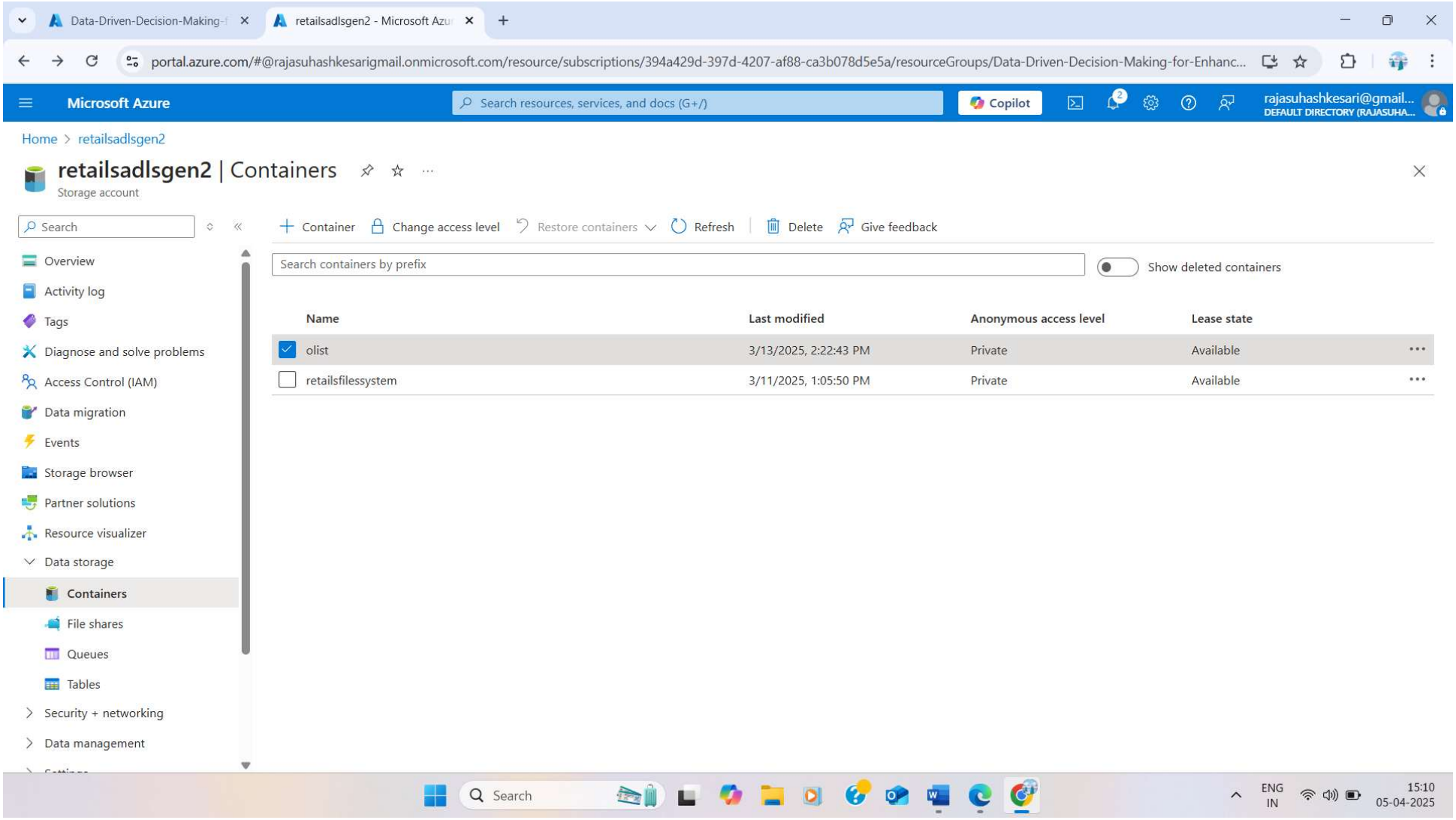
Step 1:-Created Resource Group with name ‘Data-Driven-Decision-Making-for-Enhanced-Online-Retail-Experience-and-Increased-Sales’



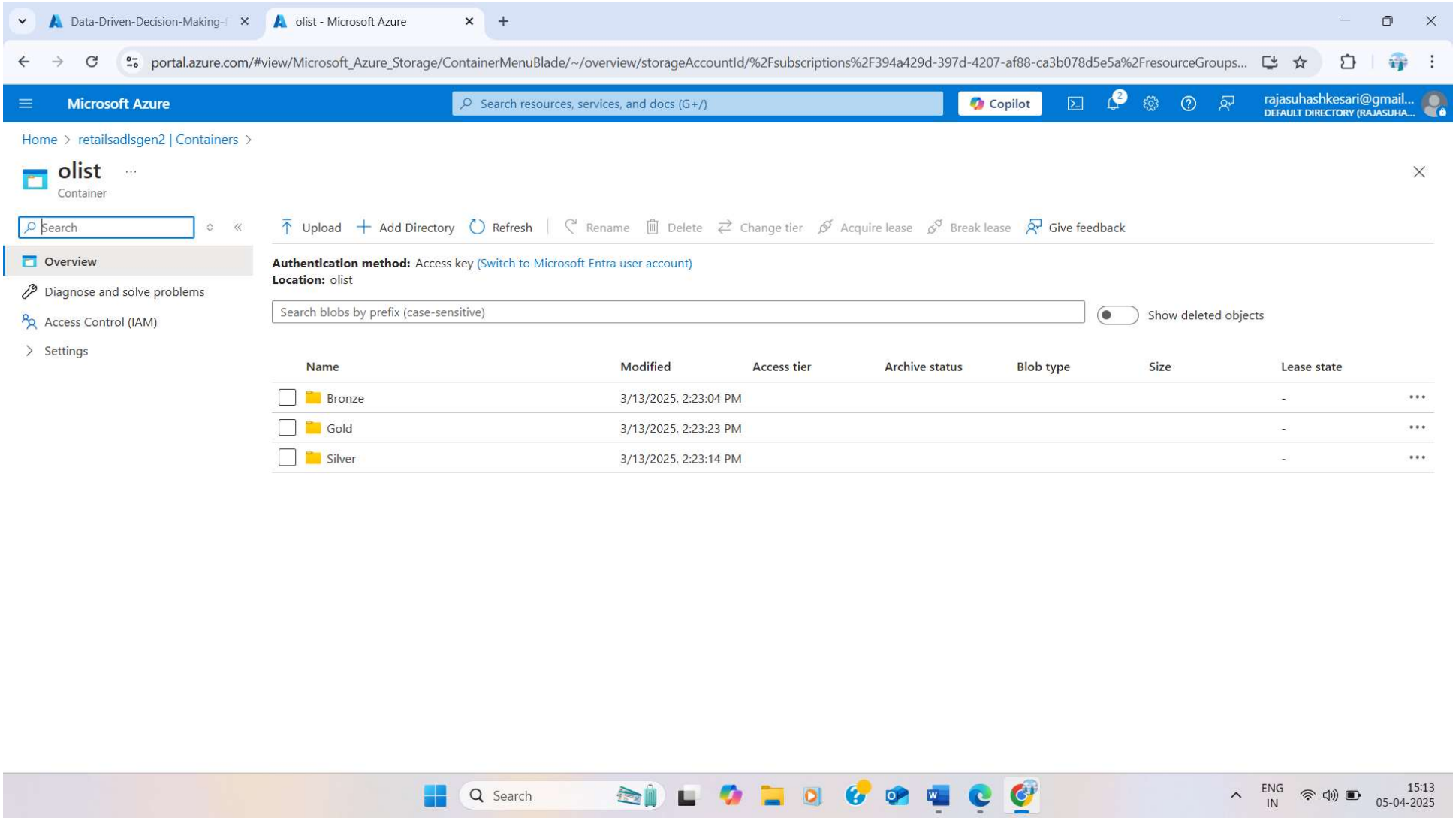
Step 2:-Created Storage Account with `*retailsadlsgen2*`name.
I created Data Lake Storage Gen2.



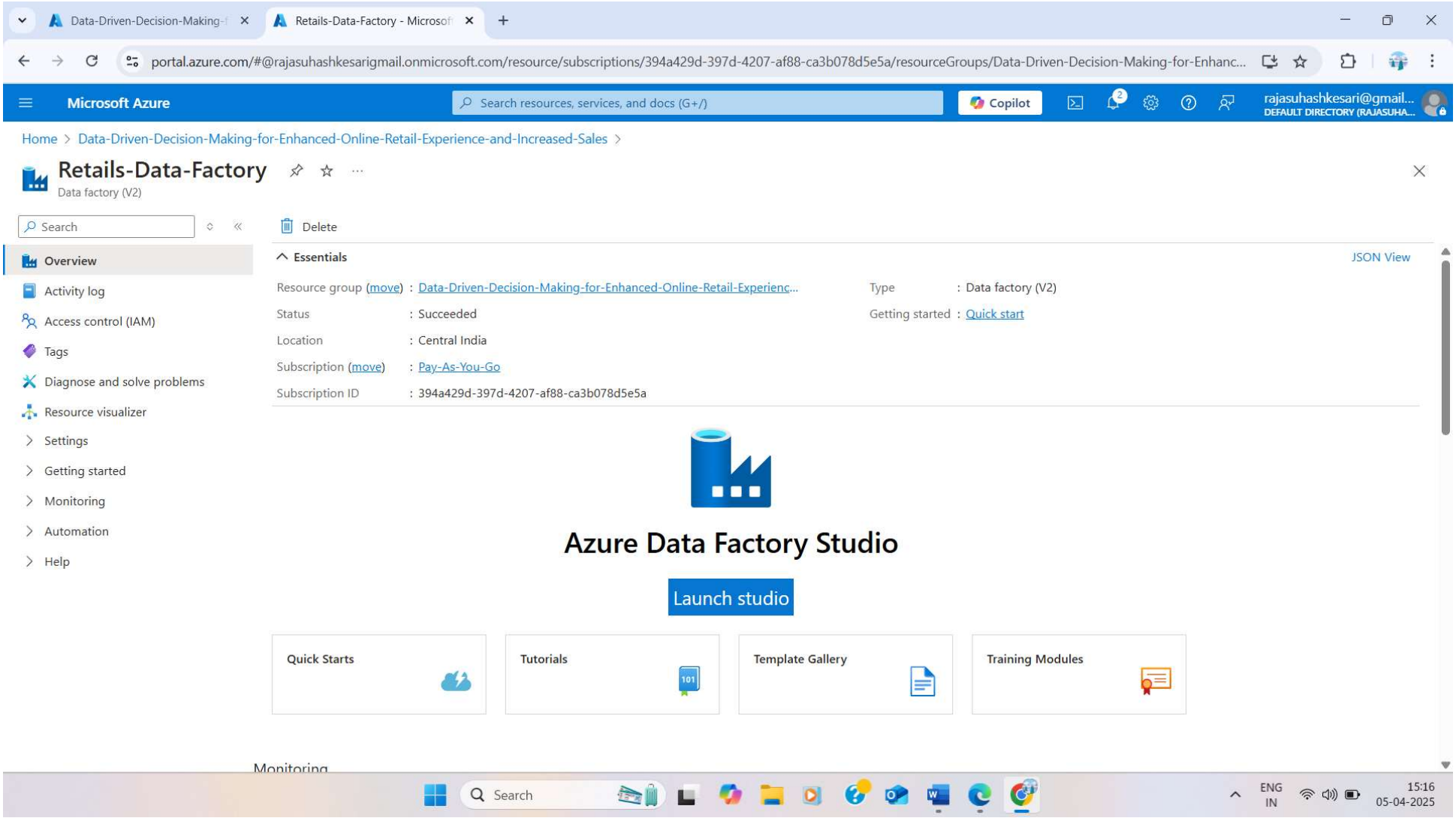
Step 3:-Created a Container ‘olist’ is the name of it.



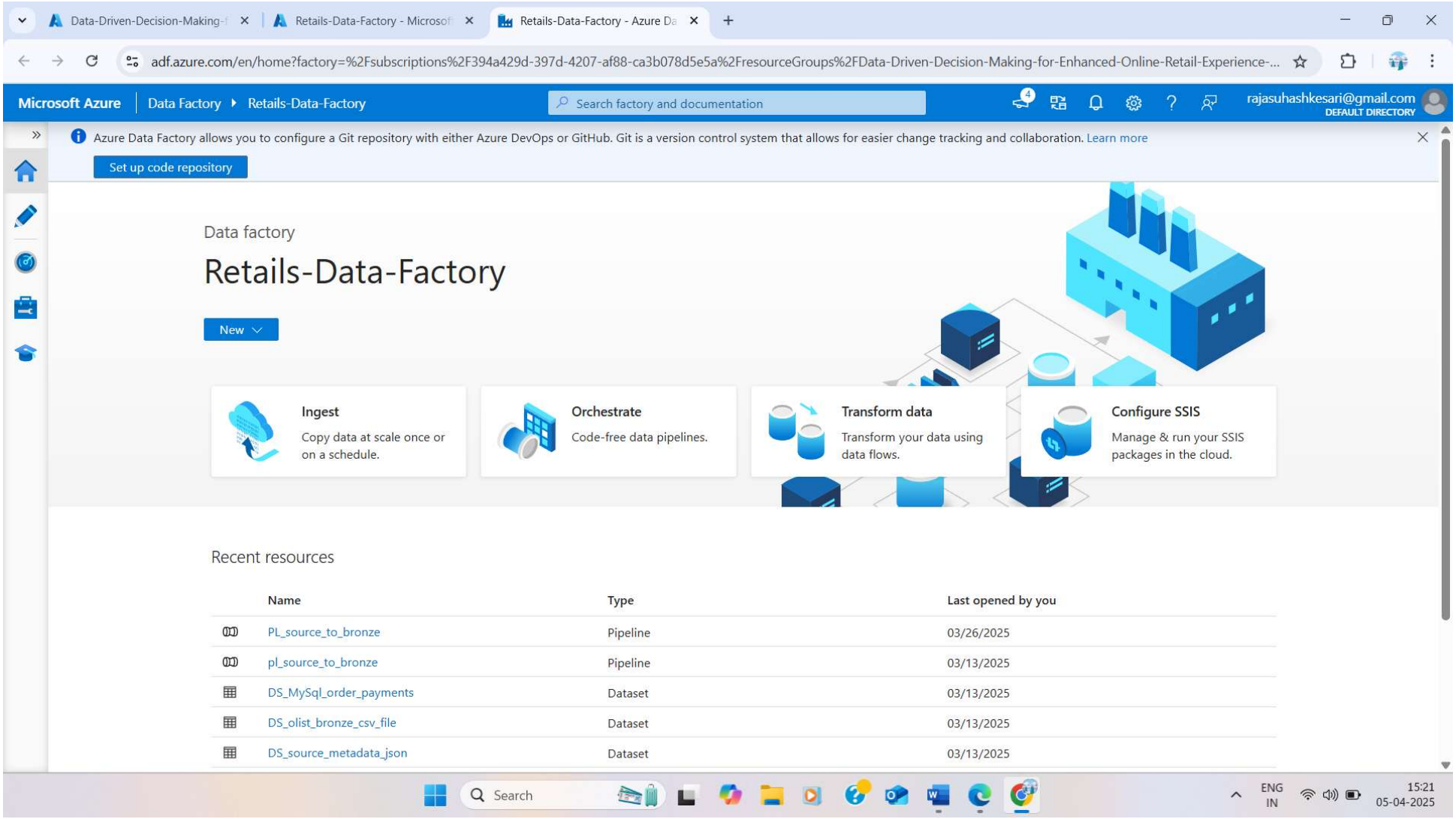
Step 4:- In ‘olist’ container I created three folders that are Bronze, Silver, Gold because we are following medallion architecture.



Step 5:- Created Azure Data Factory with name ‘Retails-Data-Factory’.



Step 6:- After Launch Azure Data Factory Studio we see like this.



Data ingestion and Load to Bronze folder in ADLS Gen2 using ADF:-

--Before ingestion we need to focus on the what is required to ingest.

1. Pipeline for which will extract data from sources and load to destination.

2. Datasets to store the metadata of the source files and to point them.

3. Linked Services to maintain the connection between ADF and sources and destinations

Creating Linked services :-

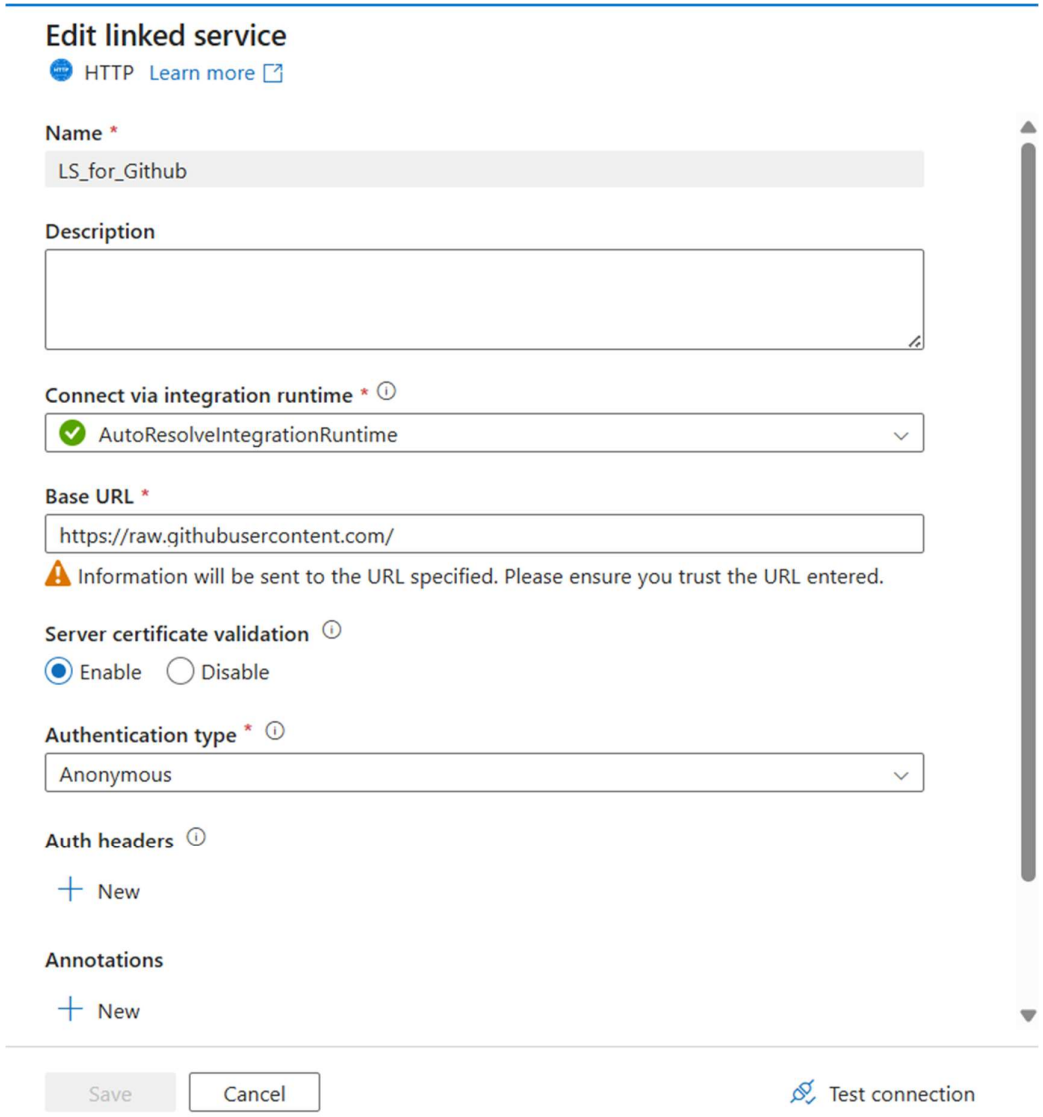
As our data required for bronze layer is in the **Github** and **MySQL database** and we have load the data to **'Bronze'** Folder in **ADLS Gen2**

So, we required three linked services:-

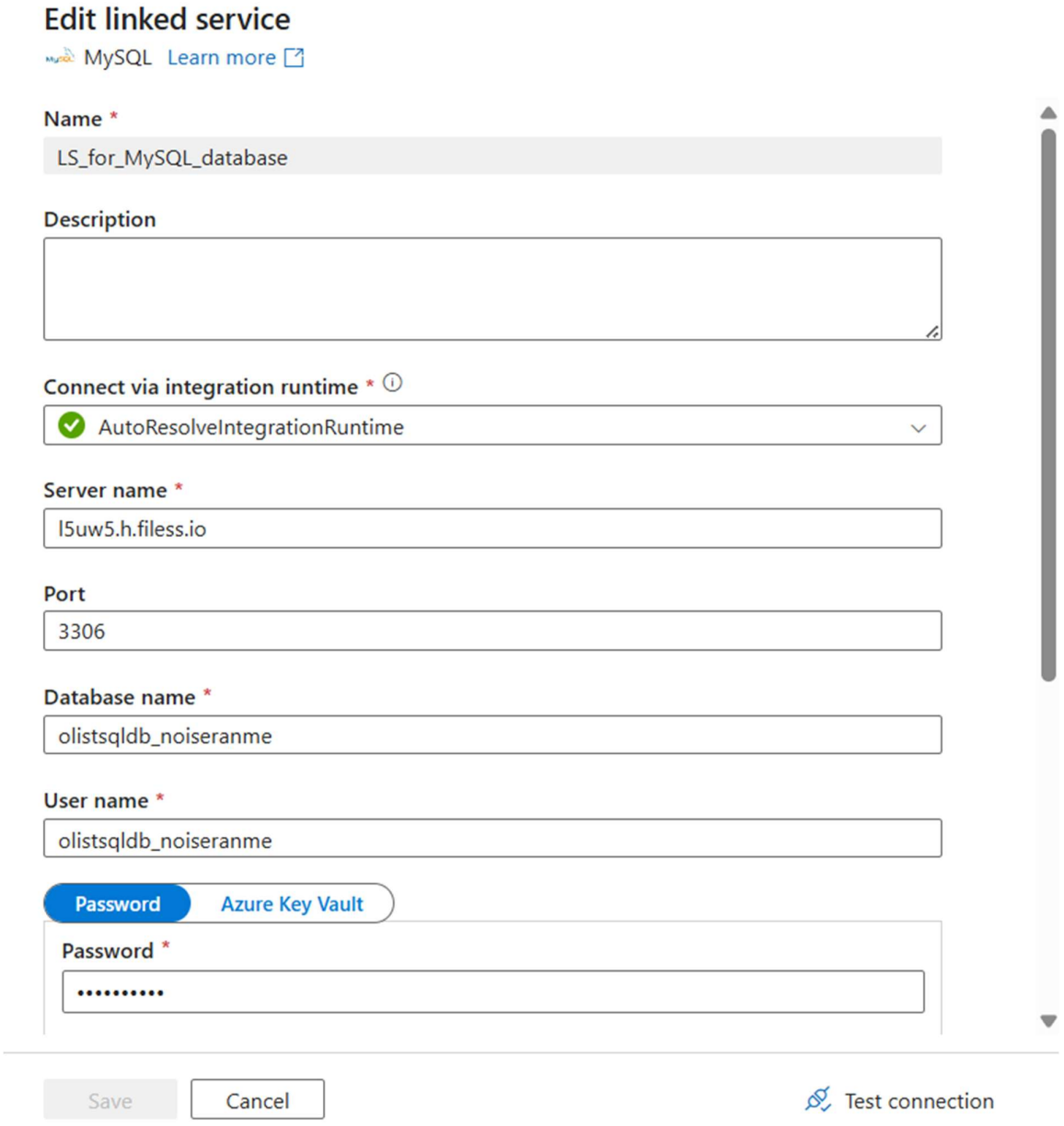
1. **Http** Linked Service to connect Github.
2. **MySQL** Linked Service to Connect MySQL database.
3. **Azure Data Lake Storage Gen2** Linked service to connected to 'Bronze' folder

I created three linked Services

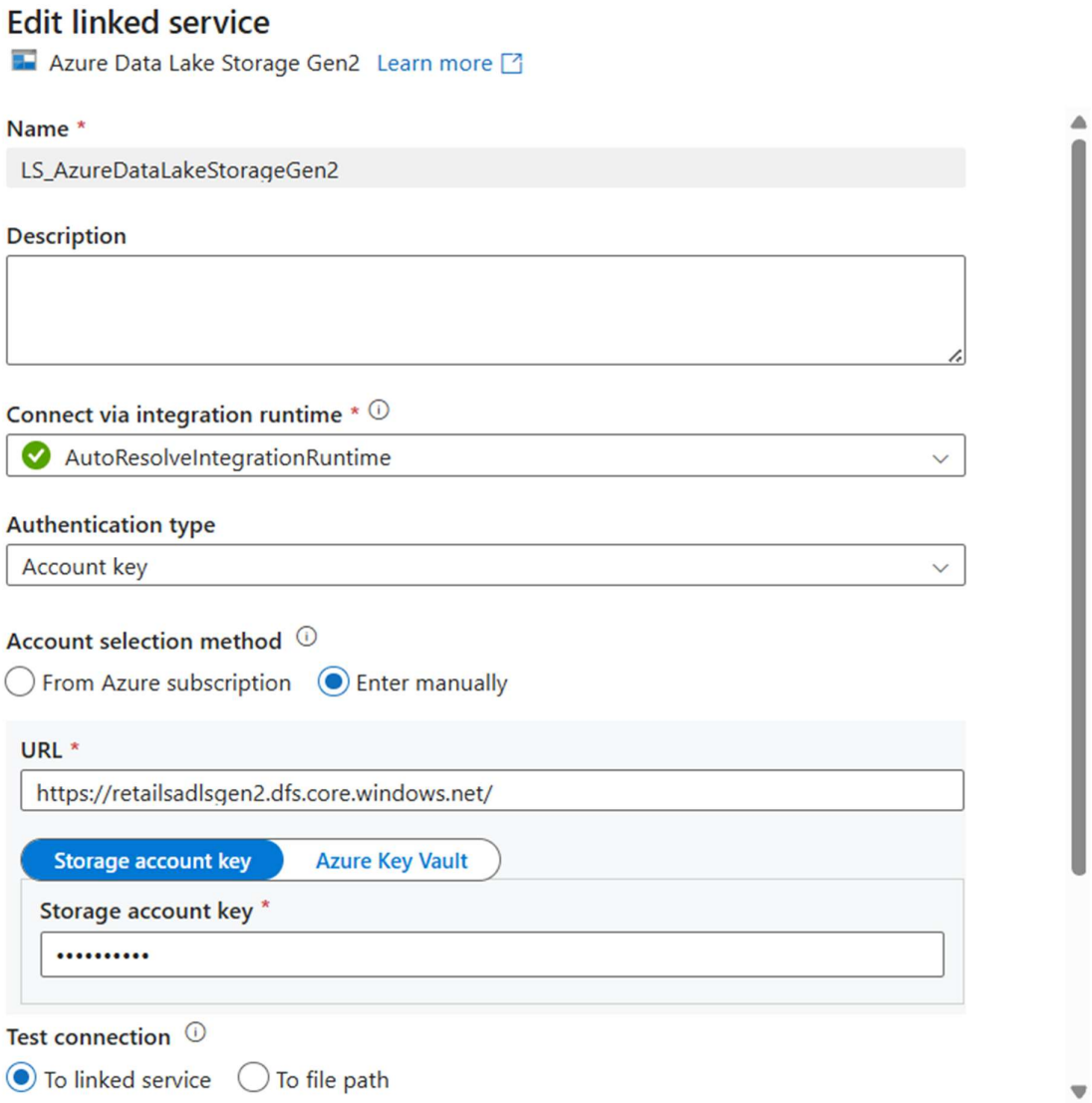
❖Created http linked service and LS_for_Github as its name



❖Created MySQL linked service and LS_for_MySQL_database as its name

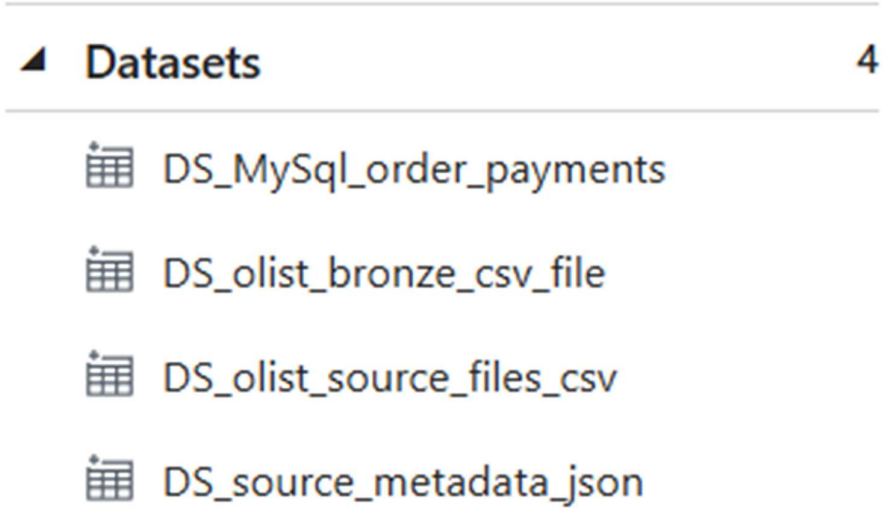


❖Created Azure Data Lake Storage Gen2 linked service and LS_AzureDataLakeStorageGen2 as its name

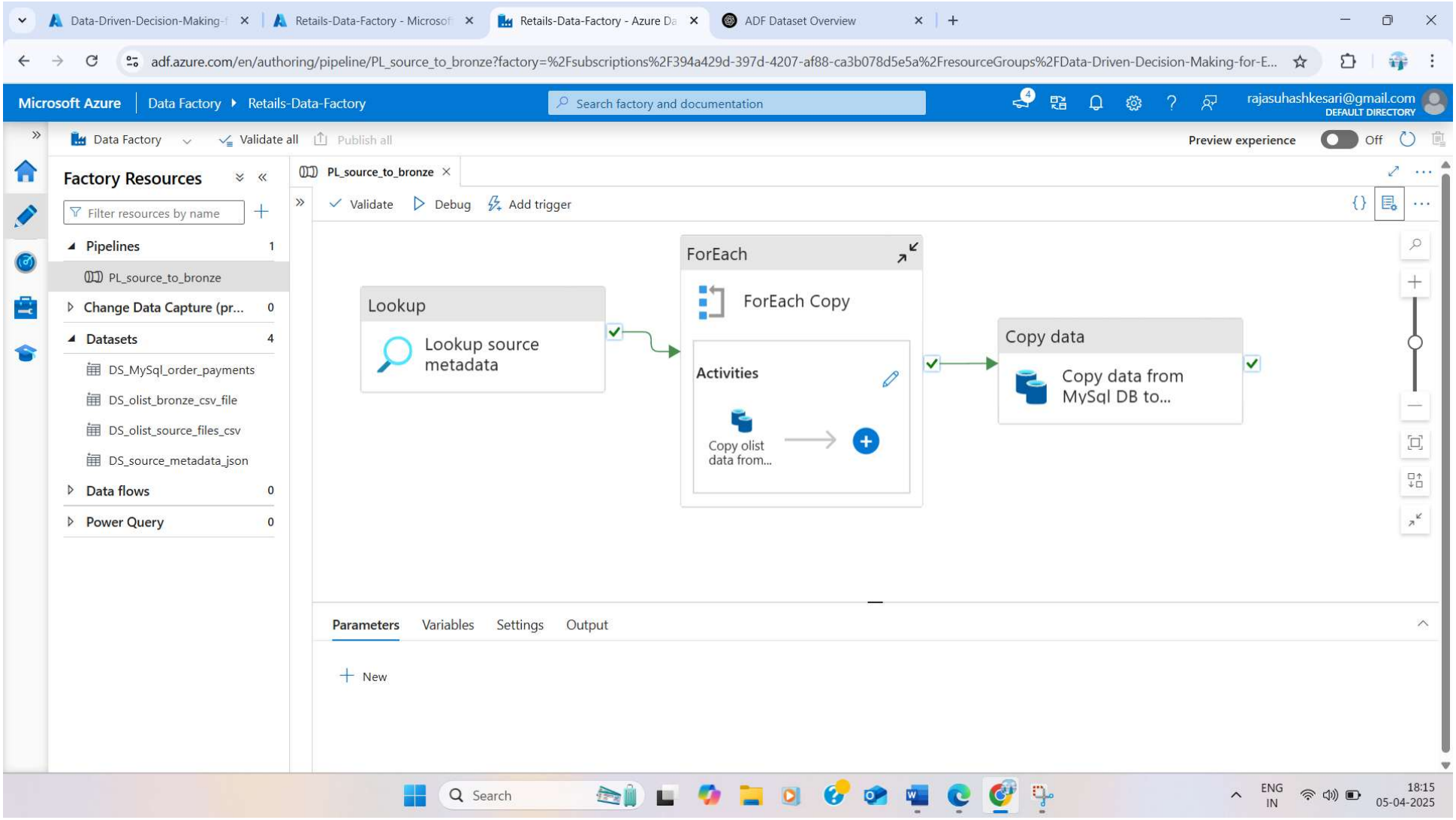


After Creating we need 4 Datasets so I created them

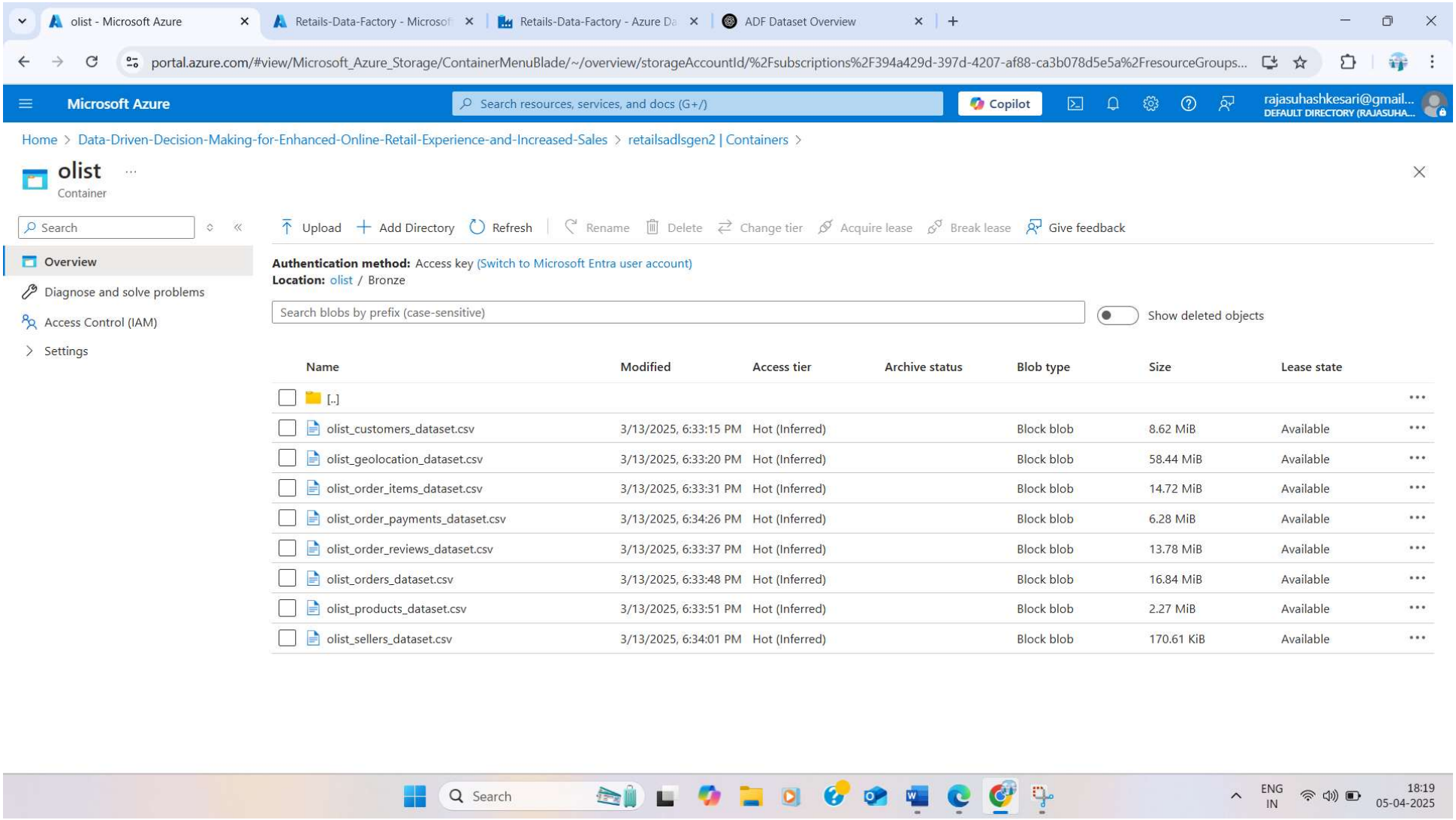
1. **DS_MySql_order_payments :-**This dataset likely refers to **order payments data stored in a MySQL database.**
2. **DS_olist_bronze_csv_file:-**This dataset is pointing to the location where the data is going to be loaded after ingestion **.**
3. **DS_olist_source_files_csv:-** This dataset is refers to the **original datasets which are csv format in the Github.**
4. **DS_source_metadata_json:-** This dataset is refers to the **metadata file in github which contains the metadata about the source files and their locations.**



--we need to a pipeline which is going to ingest the data from multiple source and load them into the bronze folder so, I created a pipline which is going to ingest the data in **github** and **Mysql table** in to **Azure Data Lake Storage Gen2 'Bronze'** folder.



After Execution of pipeline the datasets are copied from Github and MySQL database to ‘Bronze’ folder.



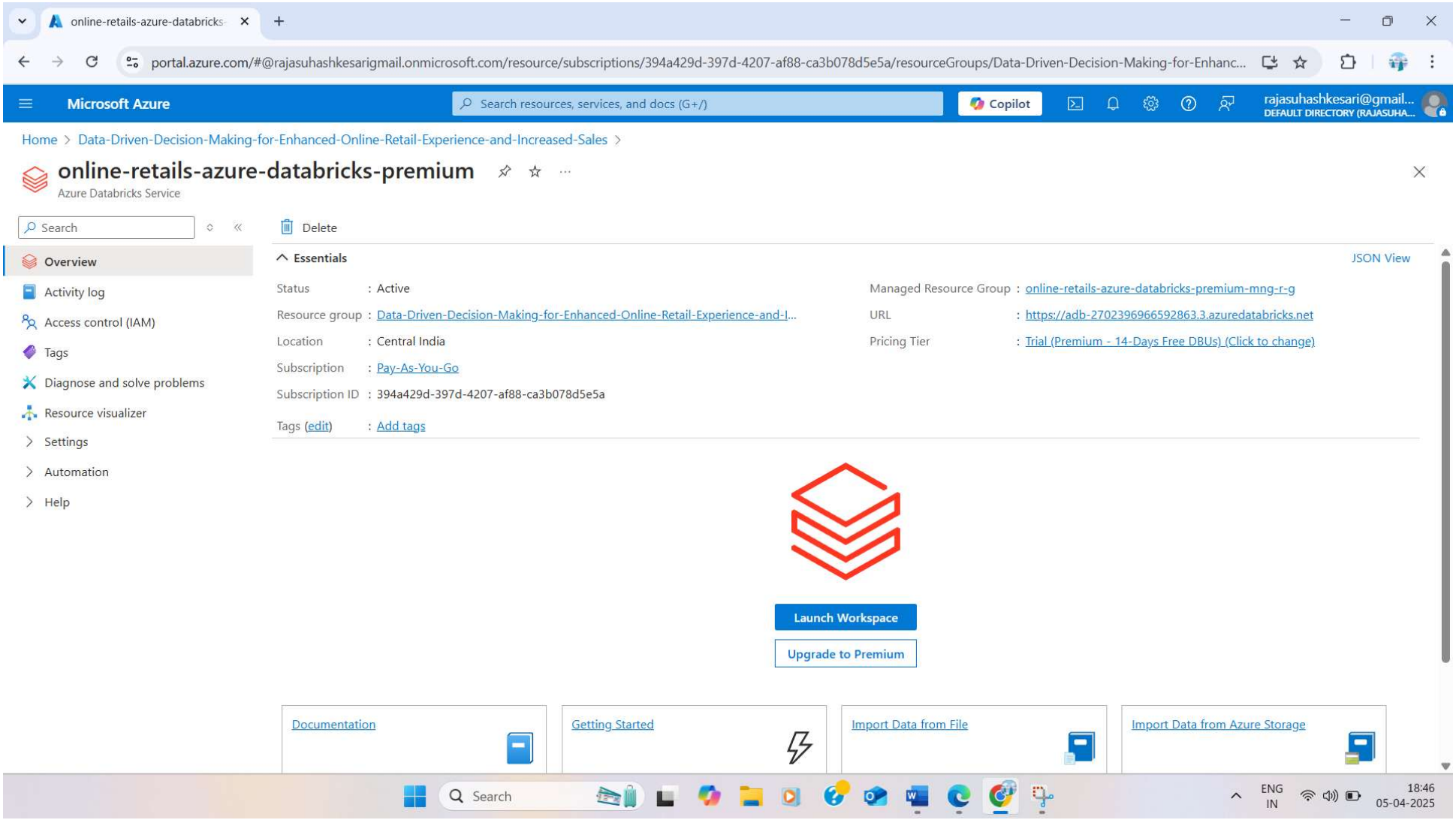
The data is successfully ingested to bronze layer using azure data factory.

Till the azure data factory role has finished in my project . now I have data in my bronze layer now I have do bring this data to Silver layer by improving the quality if the data by doing the following things

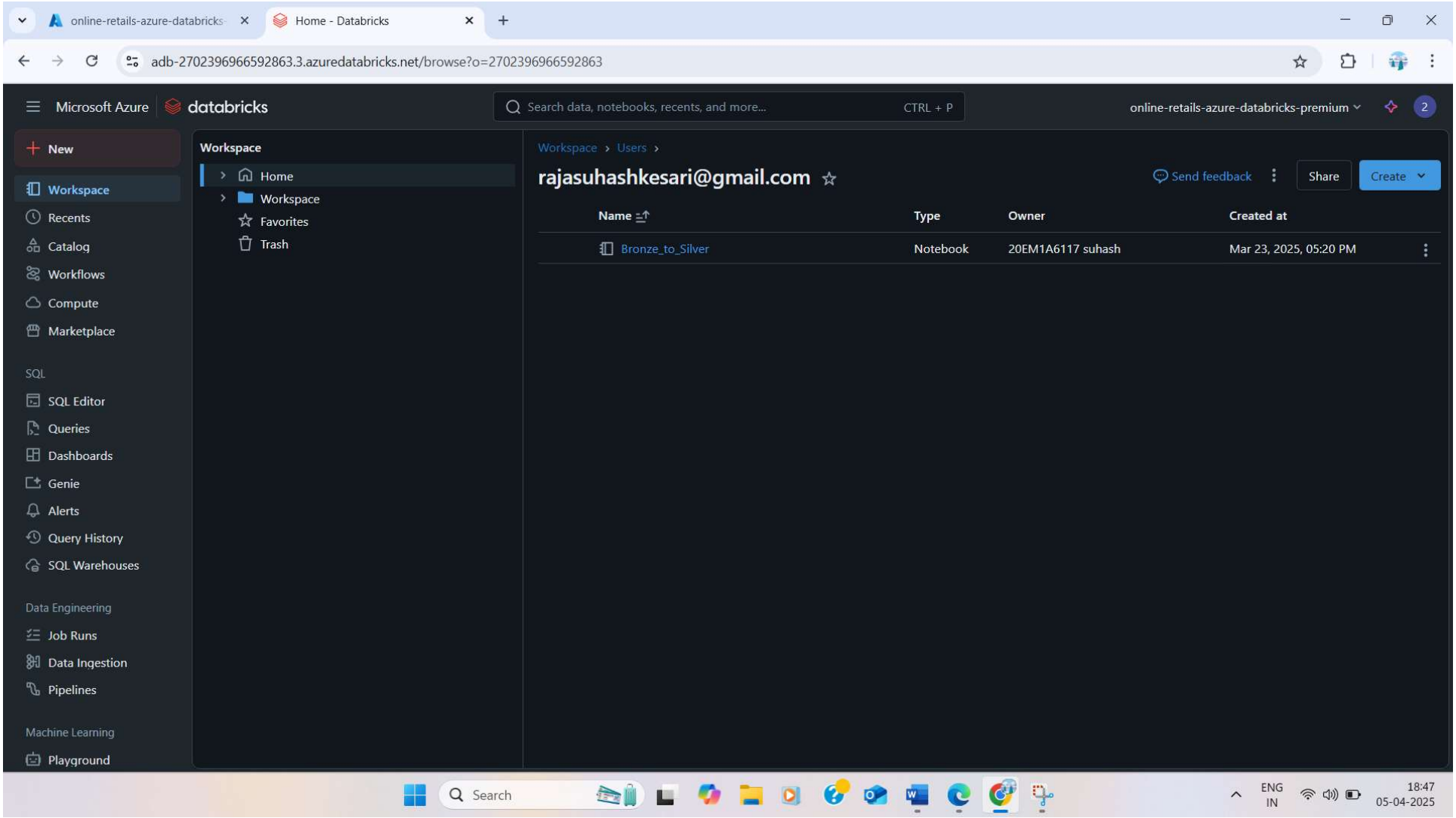
1. **Read Raw Data:**
    - Source: Bronze layer (CSV)
2. **Cleanse Data:**
    - Remove nulls, duplicates, junk rows
    - Standardize formats (dates, strings, etc.)
3. **Data Type Casting:**
    - Convert text to proper types (int, float, date)
4. **Apply Business Rules:**
    - Filter records (e.g., only completed orders)
    - Basic validations (e.g., payment > 0)
5. **Joins (Optional):**
    - Combine multiple files (e.g., orders + payments)
6. **Write to Silver:**
    - Sink: Cleaned data into Silver folder (usually in Parquet format)

As, these transformations are some more complex and it was to deal this data in the azure data factory so , I am going to used an azure databricks premium to do above things.

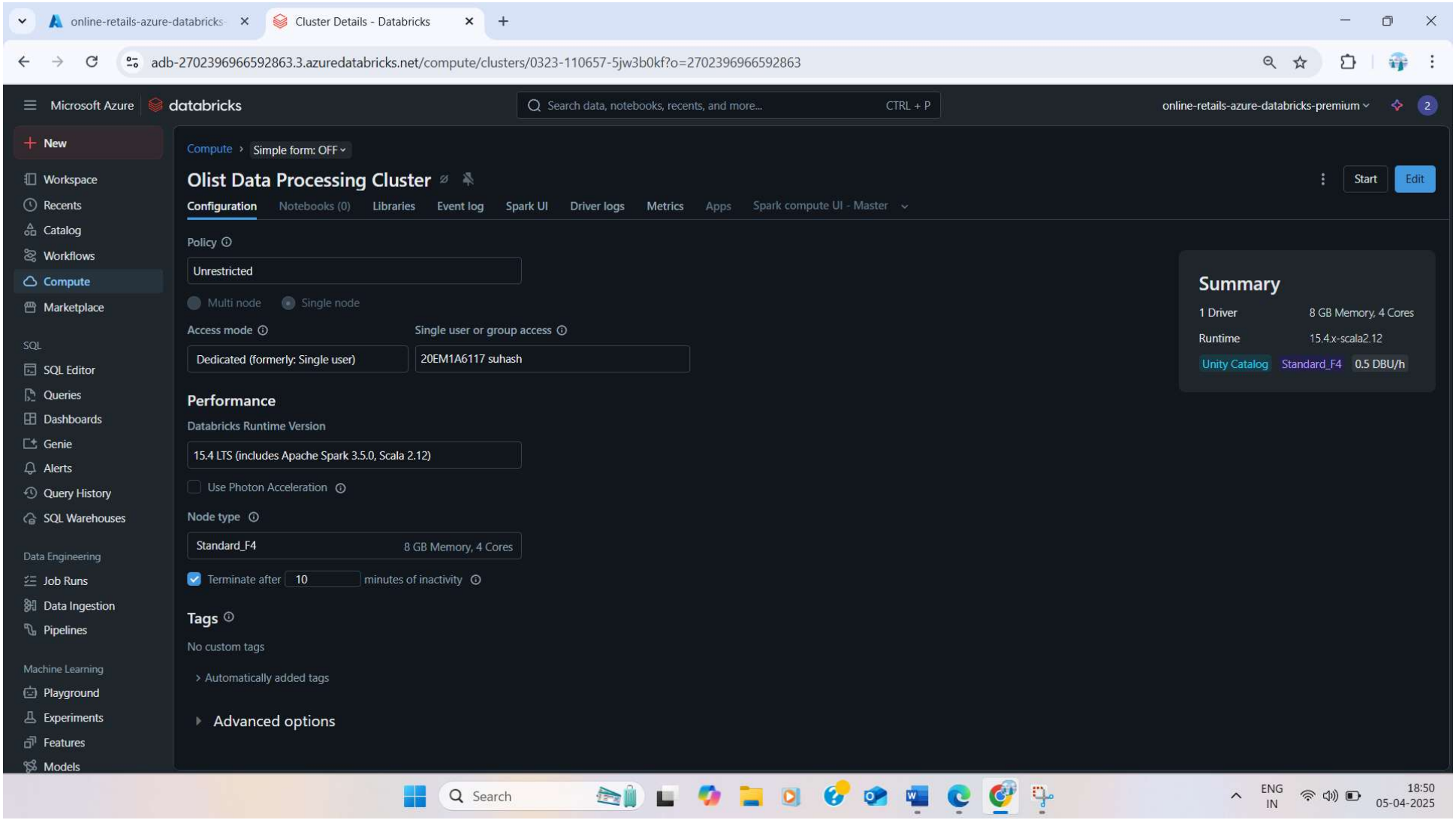
Step 7:- After getting data from sources to bronze layer I am going to use ***'Azure Databricks Premium'*** and named it as **'online-retails-azure-databricks-premium'** to data from bronze to silver by improving quality.

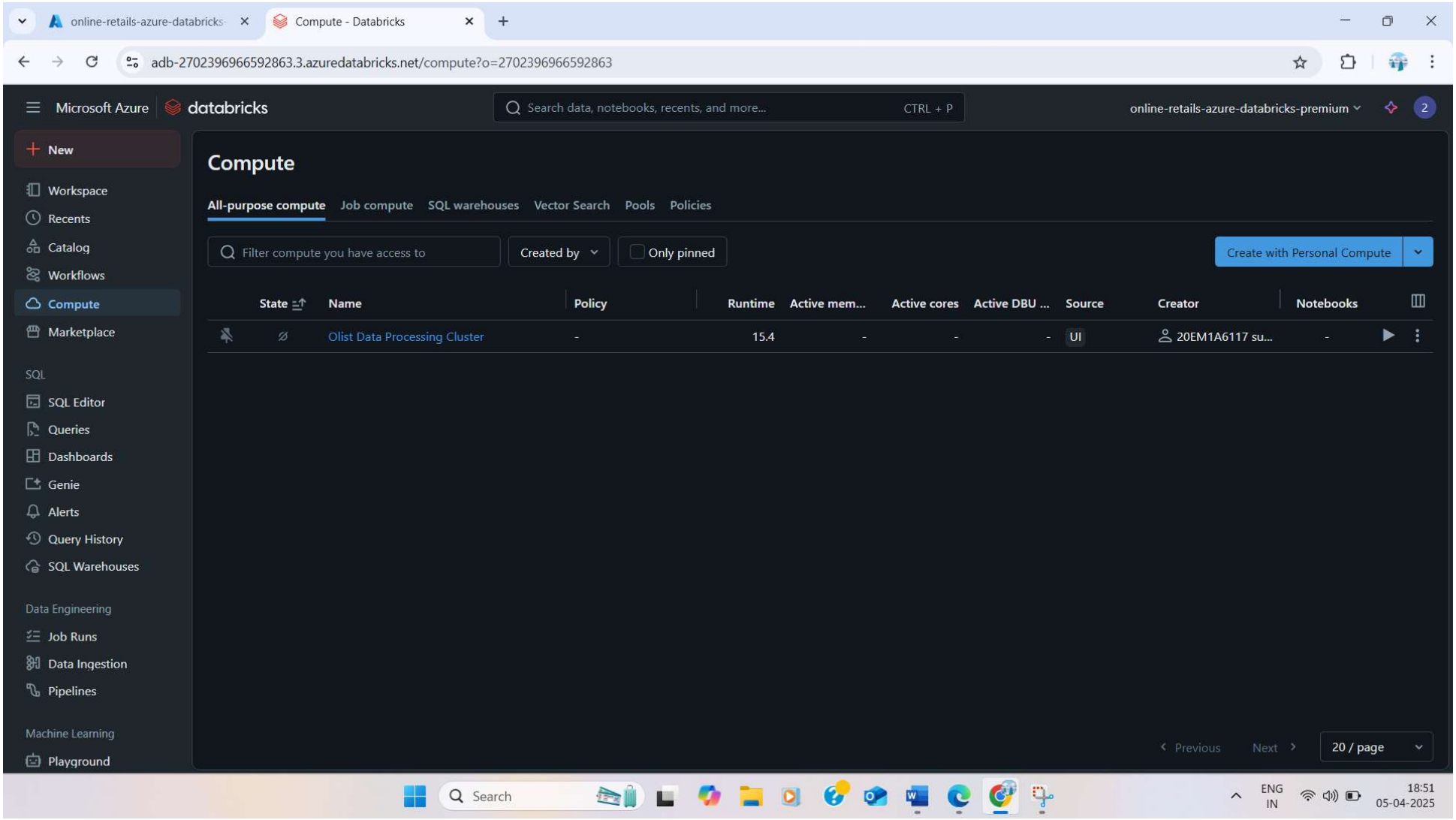


This is my Azure Databricks workspace.



To execute our ‘Notebooks’ we need compute .so, I have gone to compute section and I have created one cluster with the following requiremnts.



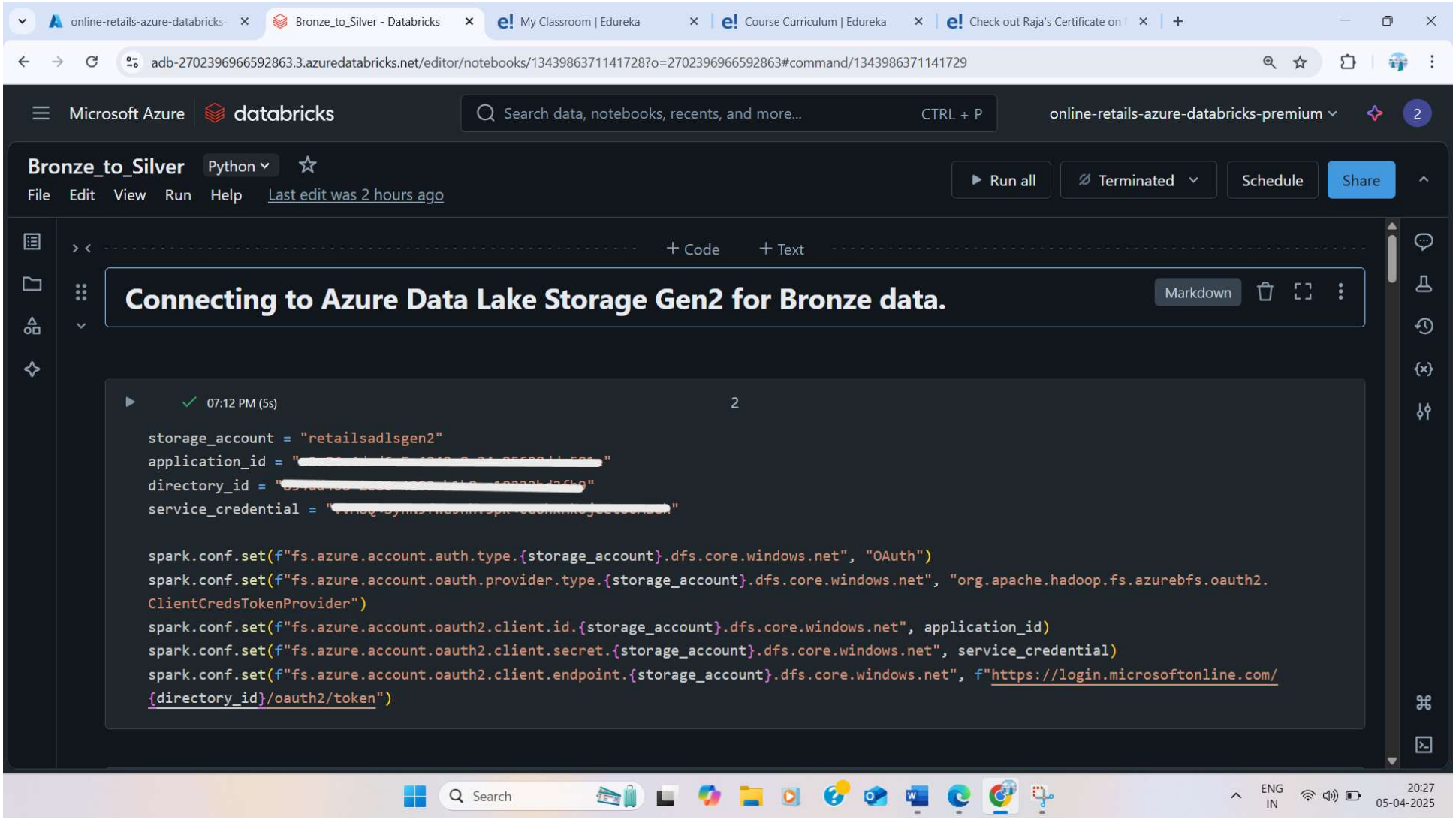


This is my cluster I was created.

I created Databricks Notebook which going to bring data from Bronze to Silver :-

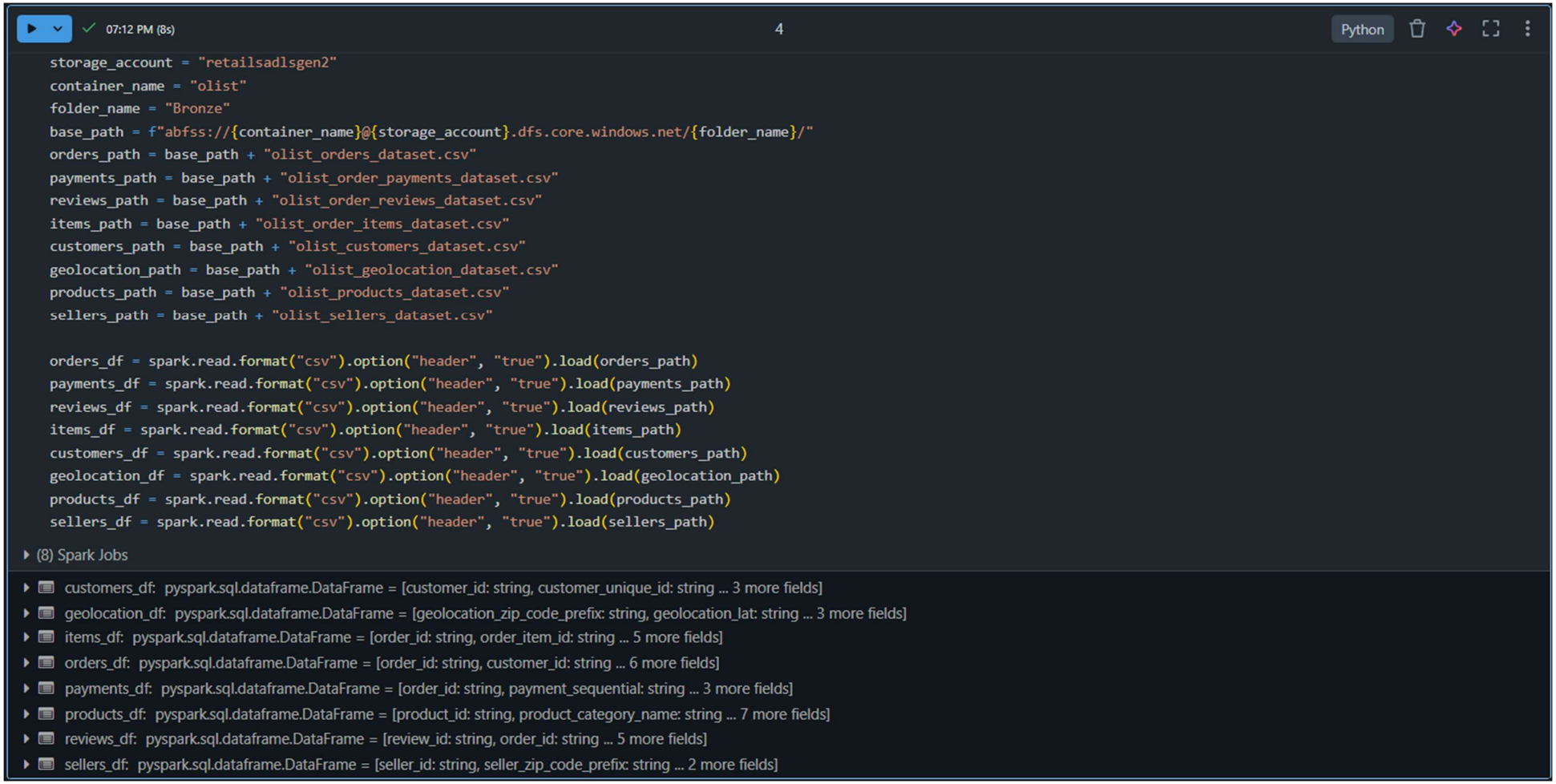
These are pyspark codes I was written in Notebook with explaination:-

Connecting to Azure Data Lake Storage Gen2 for Bronze data.

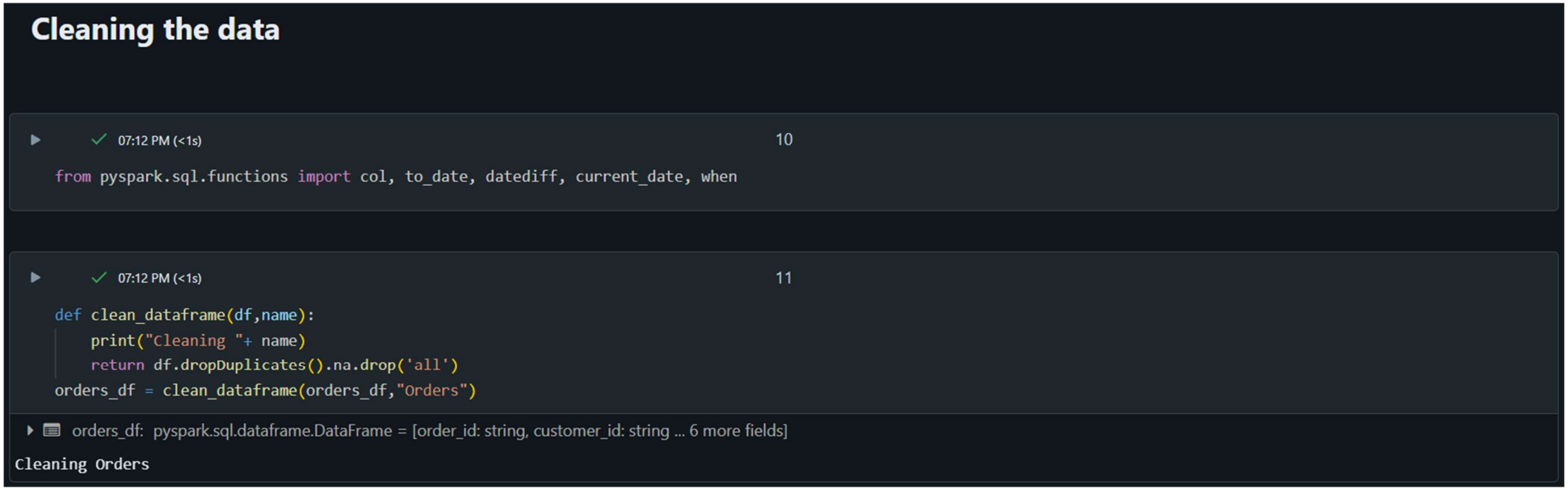


The above code sets the configurations to connect ADLS Gen2 . when I want to connect to storage account the databricks is going to connect with following above configurations. As the above some of the credentials were sensitive so I am masking them.

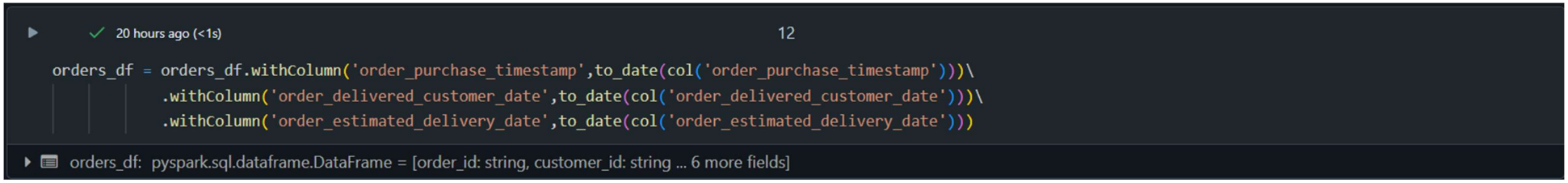
This is the pyspark will read the csv files from ADLS Gen2 and creates pyspark dataframes.



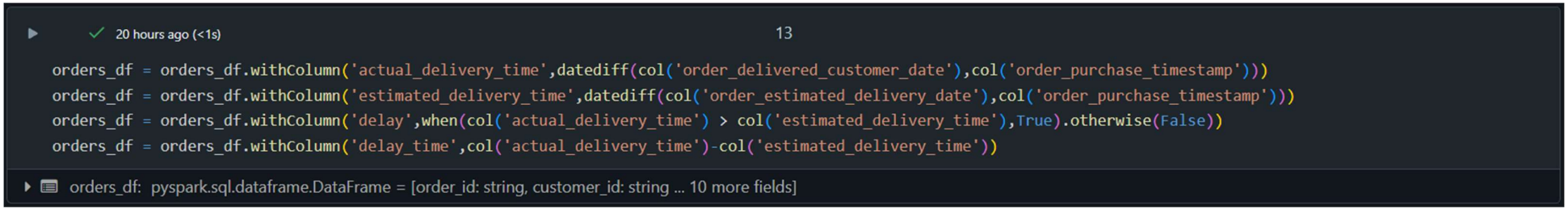
Remove duplicate and Null values :-



The Above code is going to remove the Duplicate rows and Drop rows where entire row values are null of ‘orders_df’ dataframe.



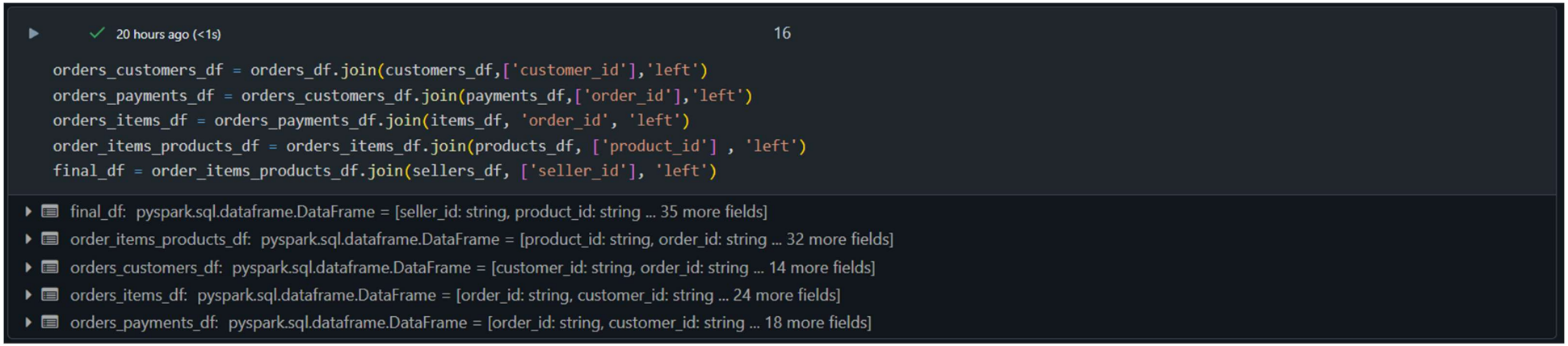
In the above code I am transforming columns timestamp to dates.



In the above pyspark code I have 4 more new columns that are

- actual_delivery_time
- estimated_delivery_time
- delay
- delay_time

Finally joined the different datasets by using the below code and data schema model



Enrichment in silver layer :-

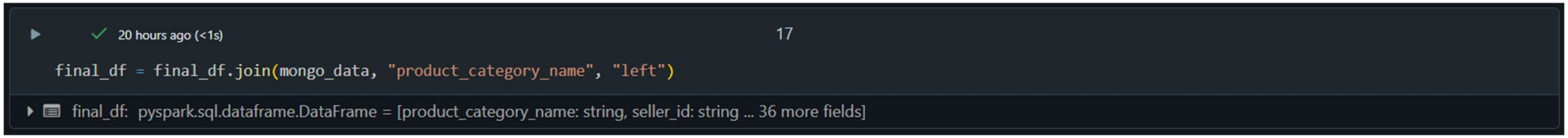
As,enrichment of data is part of the silver layer I am translating the product category names to English for this I have loaded the product_category_names from the MongoDB.

```
20 hours ago (<1s)                                                    7

# importing module
from pymongo import MongoClient

hostname = "[redacted]"
database = "[redacted]"
port = "27018"
username = "olistnosqldb_mainlytoy"
password = "[redacted]"

uri = "mongodb://" + username + ":" + password + "@" + hostname + ":" + port + "/" + database

# Connect with the portnumber and host
client = MongoClient(uri)

# Access database
mydatabase = client[database]
```

The above code connects to the MongoDB and fetches the whole database.

```
20 hours ago (3s)                                                     8

import pandas as pd
collection = mydatabase['product_category']

mongo_data = pd.DataFrame(list(collection.find()))
mongo_data = mongo_data.drop('_id',axis=1)
mongo_data = spark.createDataFrame(mongo_data)

mongo_data:  pyspark.sql.dataframe.DataFrame = [product_category_name: string, product_category_name_english: string]
```

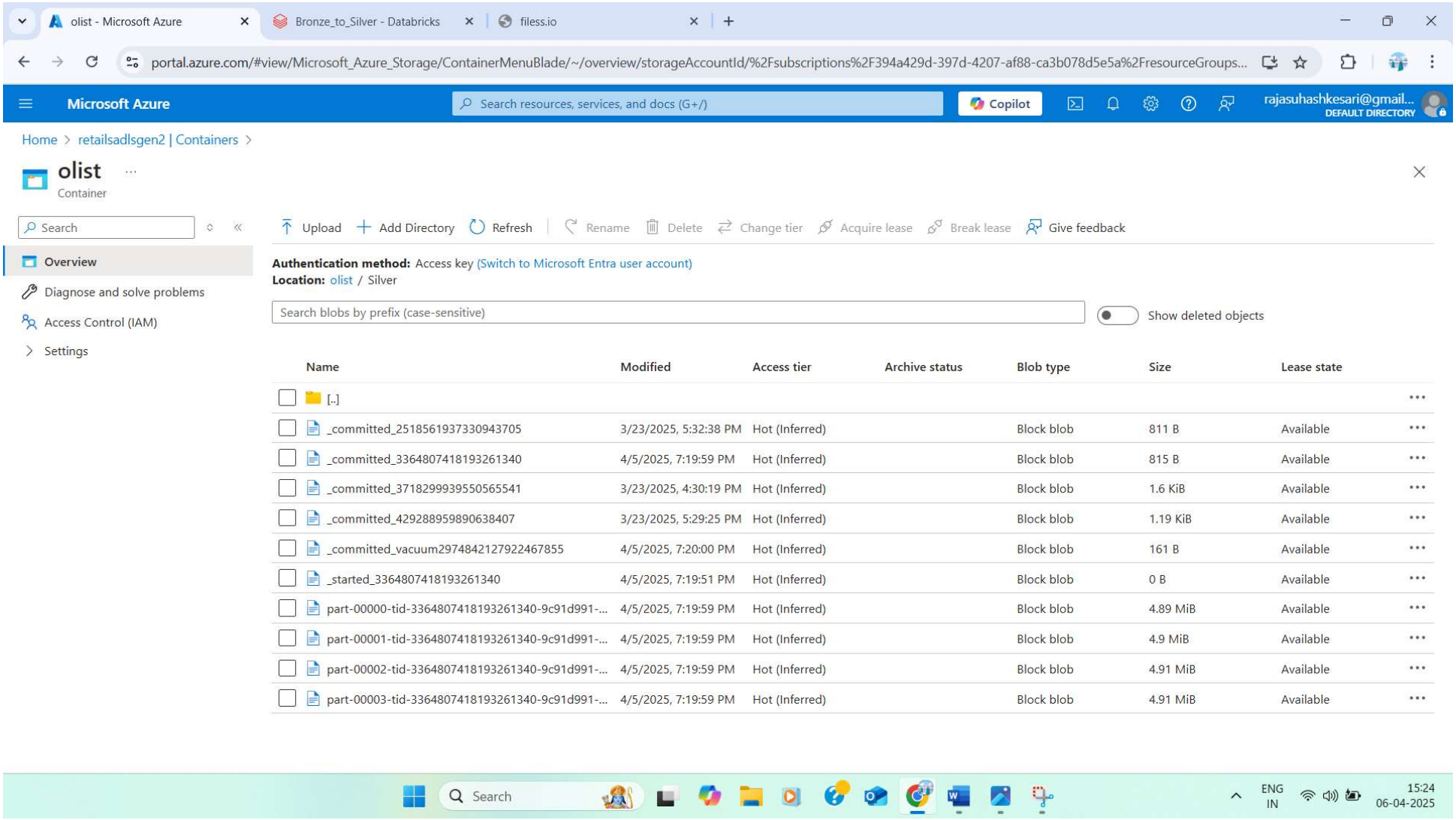
From the whole database we are fetching product category names and which in both native language and English.



Finally I joined product category names in engiles also.

Till this all my PySpark and Databricks role was finished now save the dataset to Silver folder which is in the ADLS Gen2.

This is My **‘Silver’** folder



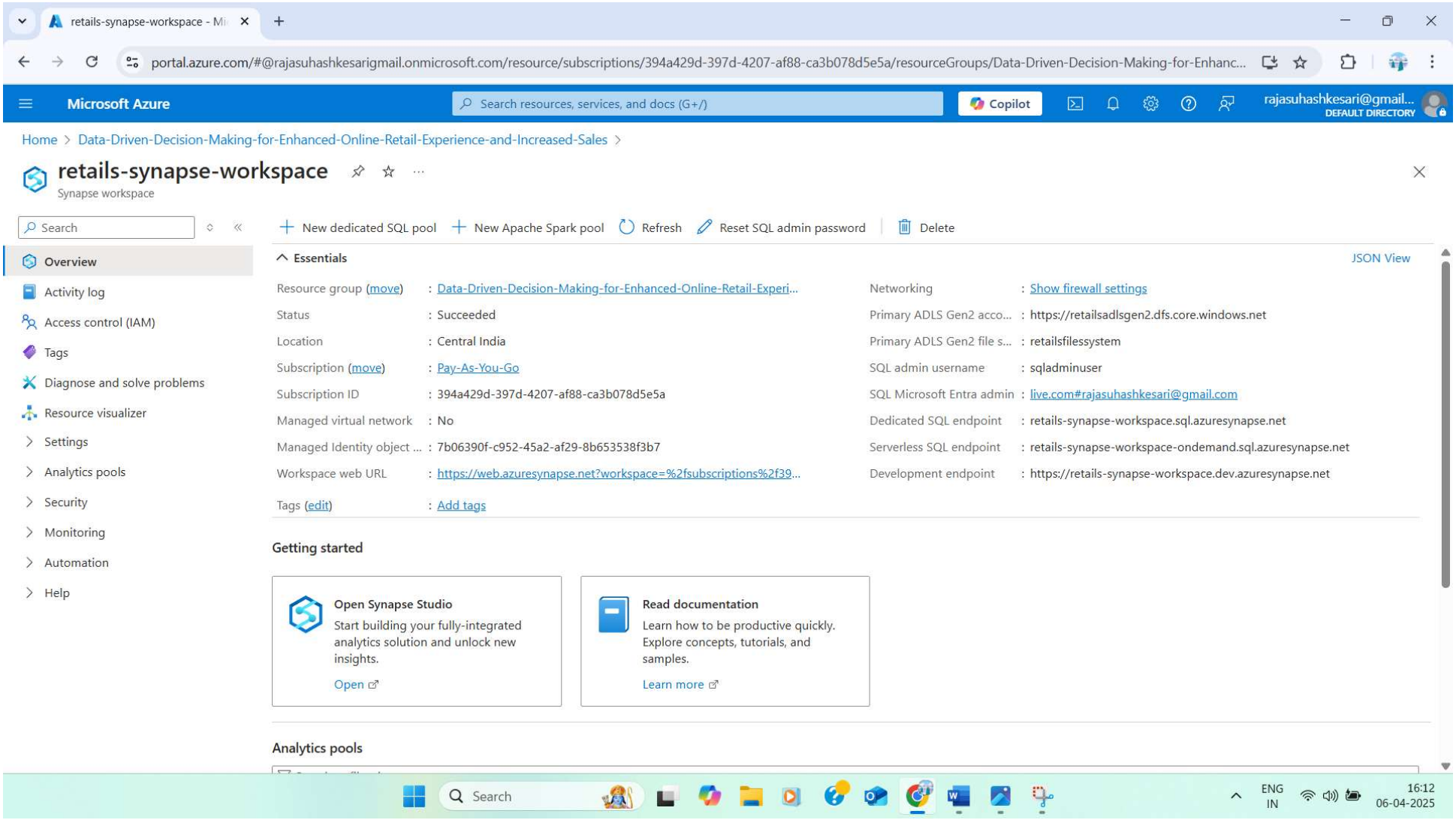
As I was used the cluster have cores so I got 4 part files, as I was executed same code multiple time so I have multiple committed files.

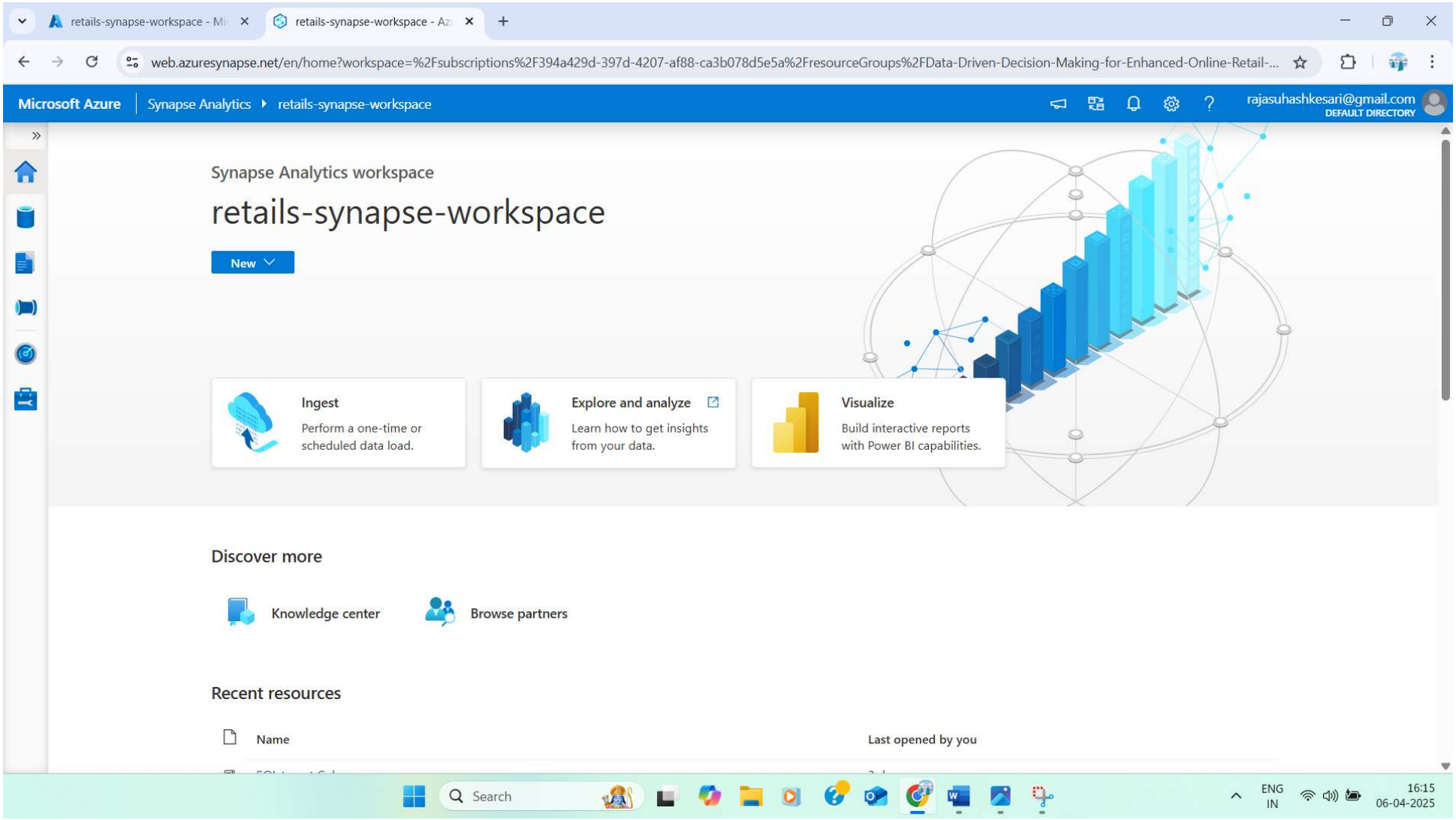
Silver to Gold Part:-

Now, we have pure data in our silver layer we will copy same data to Gold layer and then we will make different types of datasets from the Gold dataset for serving.

I have used used **‘Azure Synapse Analytics’** to bring data to gold layer for serving.

I was created an “**Azure Synapse Analytics Studio**” its names is retails-synapse-workspace .


retails-synapse-workspace - Mi
portal.azure.com/#@rajasuhashkesarigmail.onmicrosoft.com/resource/subscriptions/394a429d-397d-4207-af88-ca3b078d5e5a/resourceGroups/Data-Driven-Decision-Making-for-Enhanc...
Microsoft Azure
Search resources, services, and docs (G+/)
Copilot
rajasuhashkesari@gmail...
DEFAULT DIRECTORY
Home > Data-Driven-Decision-Making-for-Enhanced-Online-Retail-Experience-and-Increased-Sales >
retails-synapse-workspace
Synapse workspace
Search
Overview
Activity log
Access control (IAM)
Tags
Diagnose and solve problems
Resource visualizer
Settings
Analytics pools
Security
Monitoring
Automation
Help
New dedicated SQL pool
New Apache Spark pool
Refresh
Reset SQL admin password
Delete
Essentials
JSON View
Resource group (move) : Data-Driven-Decision-Making-for-Enhanced-Online-Retail-Experi...
Status : Succeeded
Location : Central India
Subscription (move) : Pay-As-You-Go
Subscription ID : 394a429d-397d-4207-af88-ca3b078d5e5a
Managed virtual network : No
Managed Identity object ... : 7b06390f-c952-45a2-af29-8b653538f3b7
Workspace web URL : https://web.azuresynapse.net?workspace=%2fsubscriptions%2f39...
Tags (edit) : Add tags
Networking : Show firewall settings
Primary ADLS Gen2 acco... : https://retailsadlsgen2.dfs.core.windows.net
Primary ADLS Gen2 file s... : retailsfilessystem
SQL admin username : sqladminuser
SQL Microsoft Entra admin : live.com#rajasuhashkesari@gmail.com
Dedicated SQL endpoint : retails-synapse-workspace.sql.azuresynapse.net
Serverless SQL endpoint : retails-synapse-workspace-ondemand.sql.azuresynapse.net
Development endpoint : https://retails-synapse-workspace.dev.azuresynapse.net
Getting started
Open Synapse Studio
Start building your fully-integrated analytics solution and unlock new insights.
Open
Read documentation
Learn how to be productive quickly. Explore concepts, tutorials, and samples.
Learn more
Analytics pools
Search
ENG
IN
16:12
06-04-2025


retails-synapse-workspace - Mi
retails-synapse-workspace - Az
web.azuresynapse.net/en/home?workspace=%2Fsubscriptions%2F394a429d-397d-4207-af88-ca3b078d5e5a%2FresourceGroups%2FData-Driven-Decision-Making-for-Enhanced-Online-Retail-...
Microsoft Azure
Synapse Analytics
retails-synapse-workspace
rajasuhashkesari@gmail.com
DEFAULT DIRECTORY
Synapse Analytics workspace
retails-synapse-workspace
New
Ingest
Perform a one-time or scheduled data load.
Explore and analyze
Learn how to get insights from your data.
Visualize
Build interactive reports with Power BI capabilities.
Discover more
Knowledge center
Browse partners
Recent resources
Name
Last opened by you
Search
ENG
IN
16:15
06-04-2025

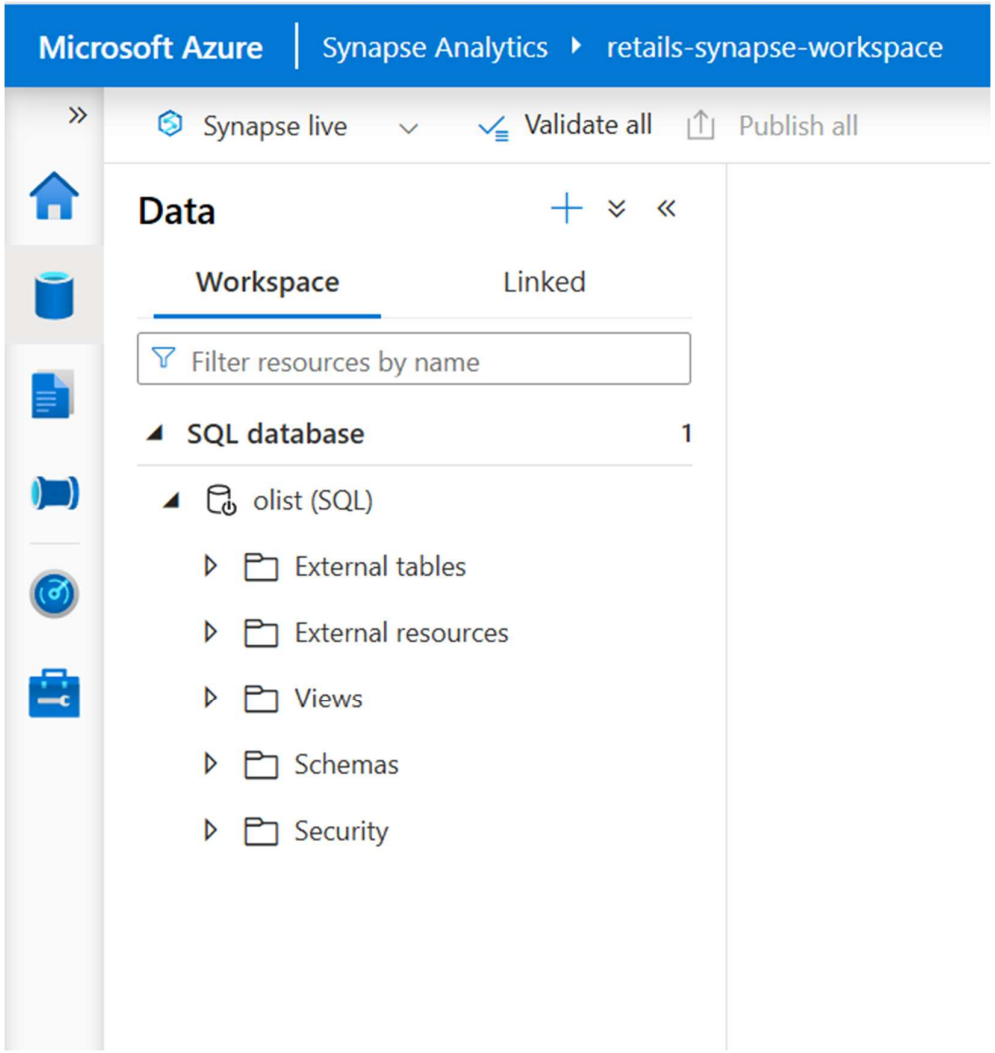
In synapse I am using serverless compute services and External tables to make gold data

Benefits of synapse :- By using Synapse we can query the data in csv,parquet using SQL query

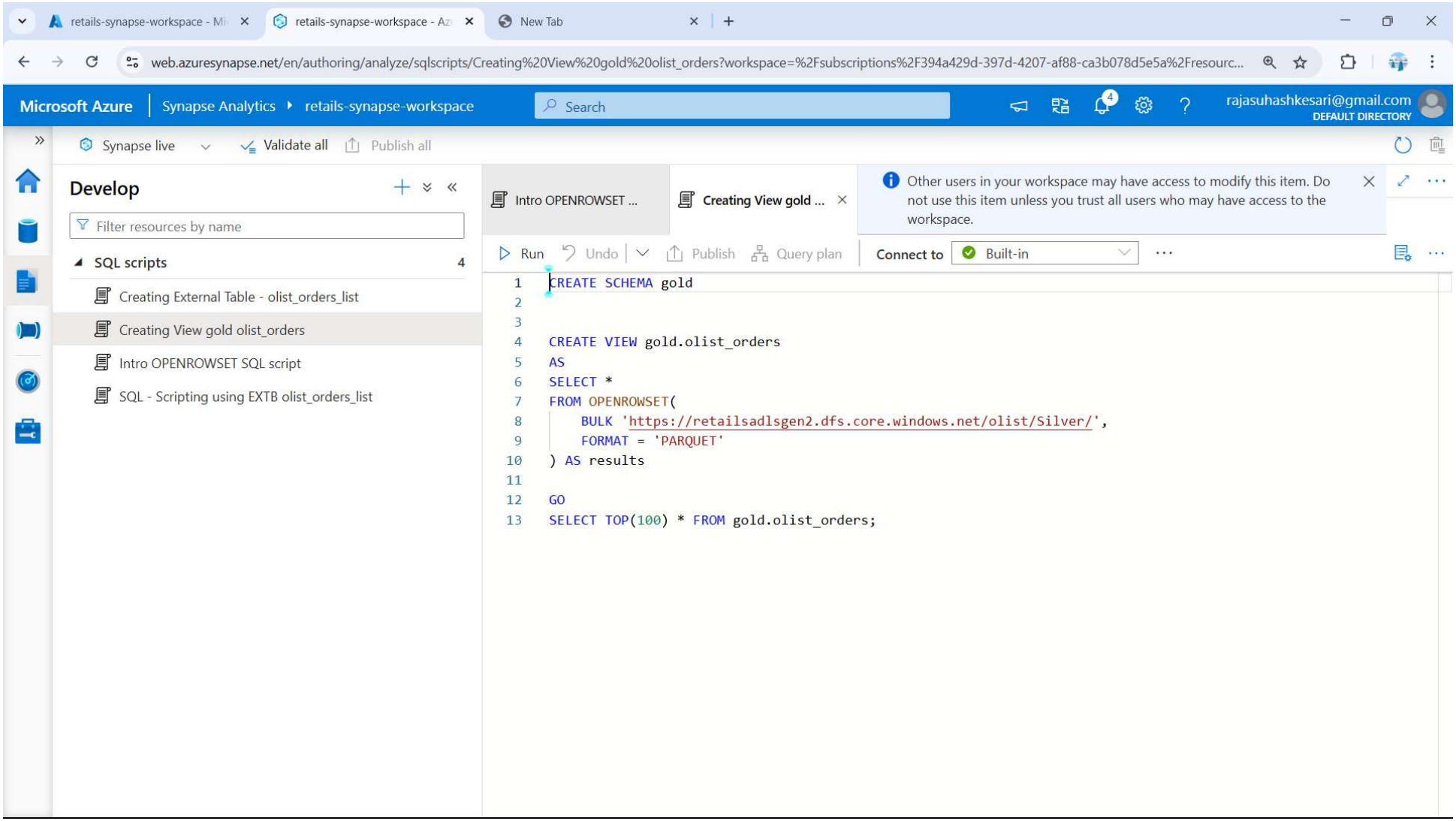
Example:-

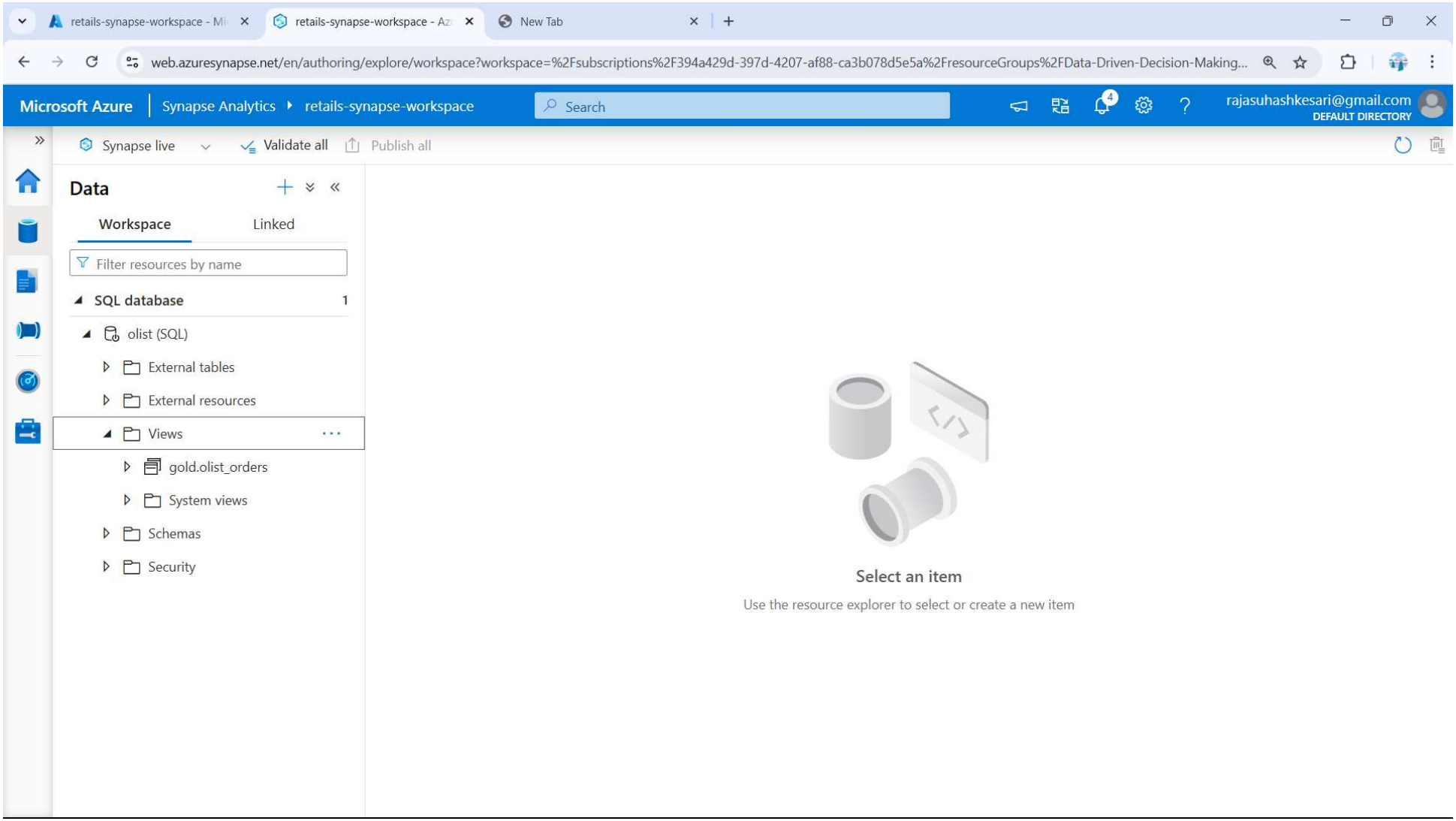
```
SELECT *
FROM OPENROWSET(
    BULK 'https://retailsadlsgen2.dfs.core.windows.net/olist/Silver/',
    FORMAT = 'PARQUET'
) AS results
```

Next, I created an Serverless Sql database name is ‘olist’



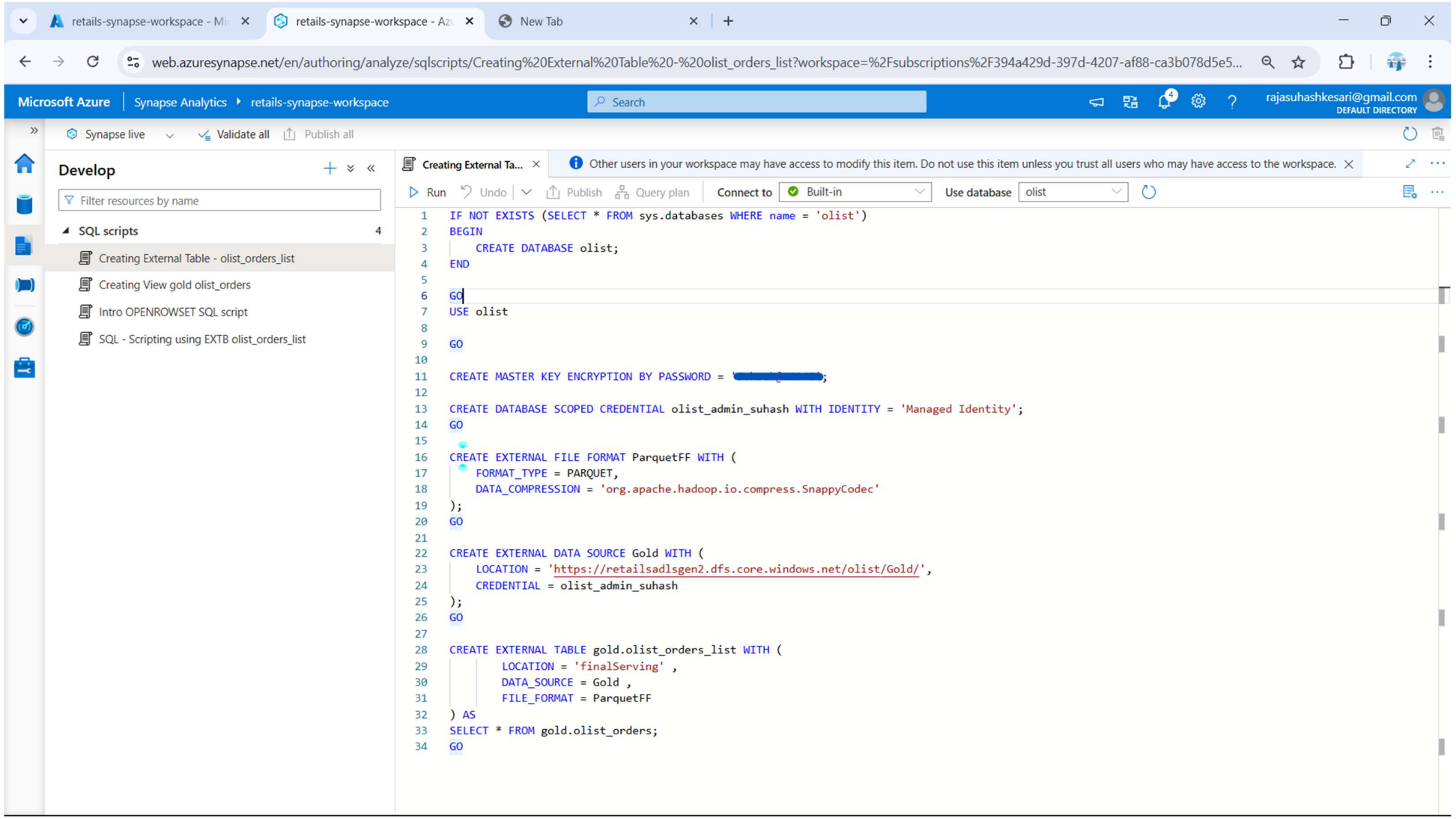
I have created a SQL VIEW which will fetch the data from the silver layer using sql query.



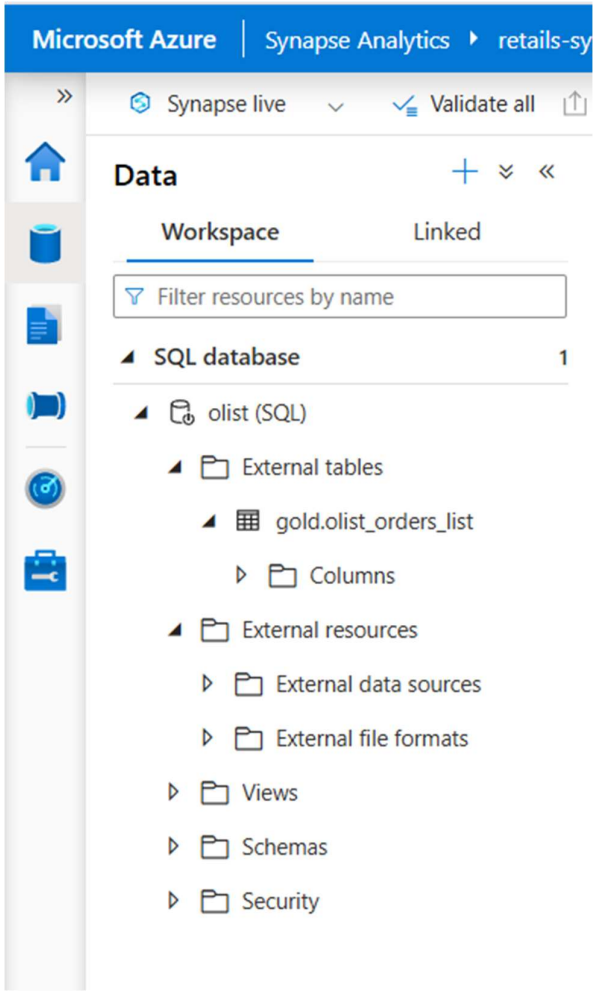

retails-synapse-workspace - Mi
retails-synapse-workspace - Az
New Tab
web.azuresynapse.net/en/authoring/explore/workspace?workspace=%2Fsubscriptions%2F394a429d-397d-4207-af88-ca3b078d5e5a%2FresourceGroups%2FData-Driven-Decision-Making...
Microsoft Azure
Synapse Analytics
retails-synapse-workspace
Search
rajasuhashkesari@gmail.com
DEFAULT DIRECTORY
Synapse live
Validate all
Publish all
Data
Workspace
Linked
Filter resources by name
SQL database
1
olist (SQL)
External tables
External resources
Views
gold.olist_orders
System views
Schemas
Security
Select an item
Use the resource explorer to select or create a new item

Creating external table in azure synapse :-

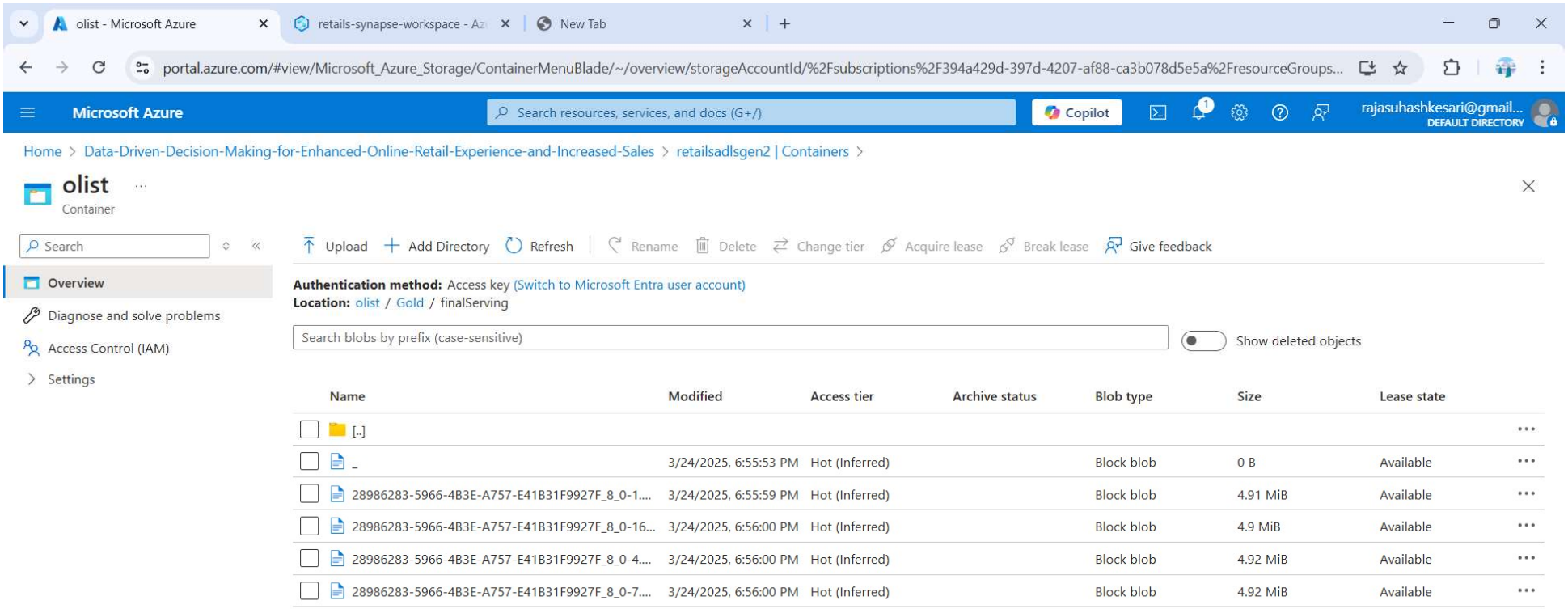
By using this sql code I created an external table .



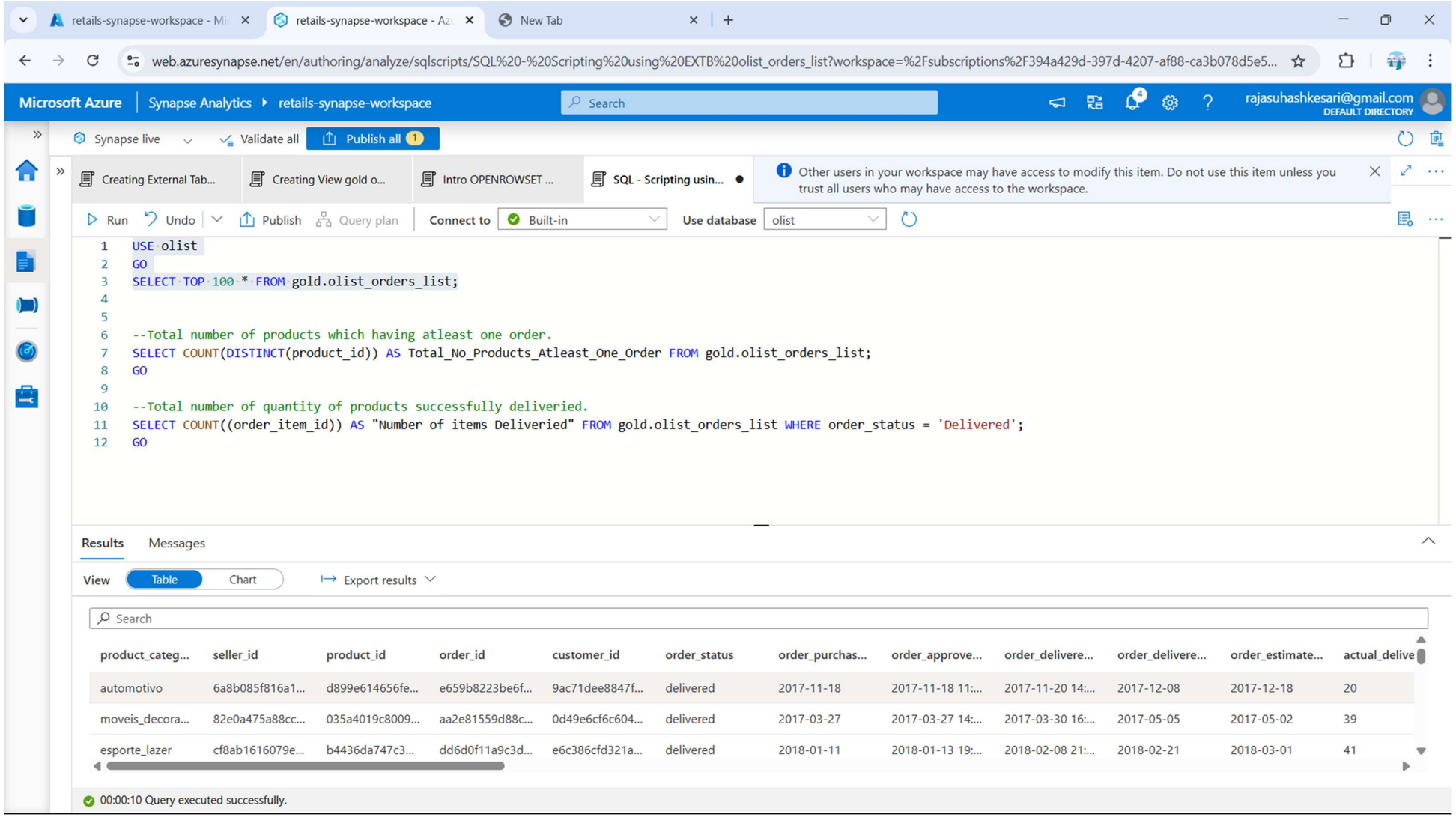
After creation of EXTERNAL TABLE I seen the table inside the olist database

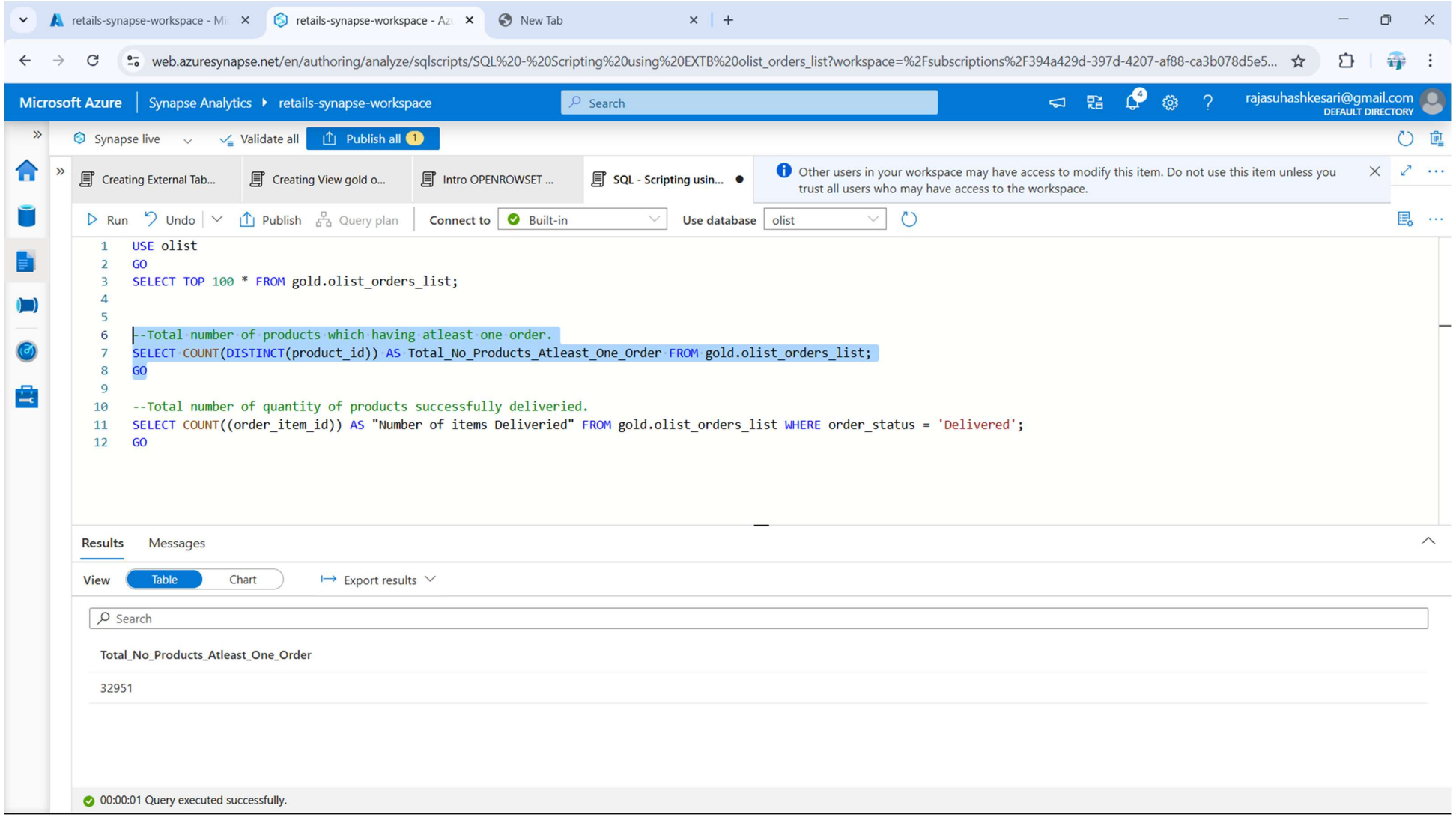


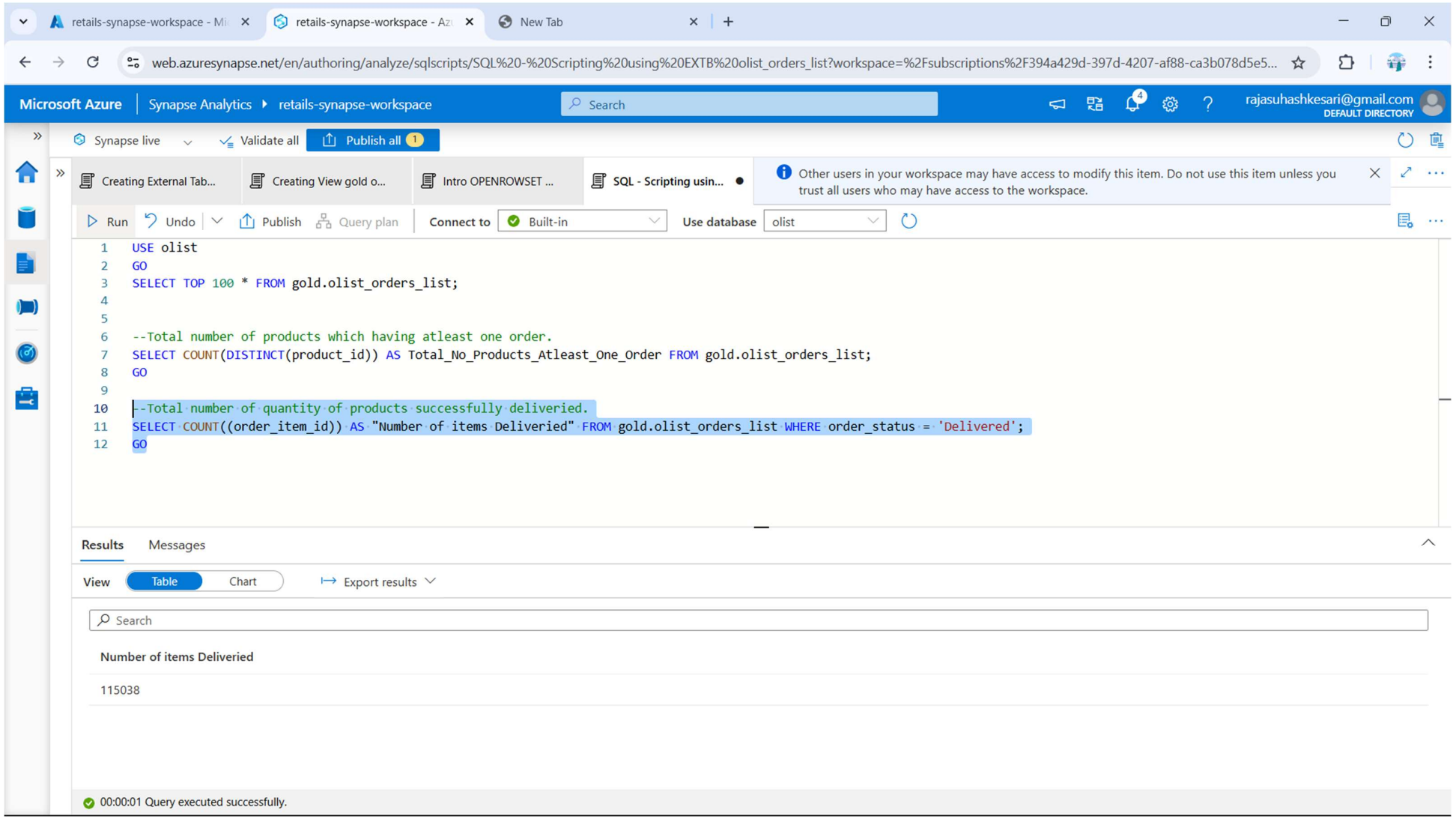
After creating external table I got data from silver to gold



Now, the data is ready for serving let us check with some sql scripts


retails-synapse-workspace - Mi
retails-synapse-workspace - Az
New Tab
web.azuresynapse.net/en/authoring/analyze/sqlscripts/SQL%20-%20Scripting%20using%20EXTB%20olist_orders_list?workspace=%2Fsubscriptions%2F394a429d-397d-4207-af88-ca3b078d5e5...
Microsoft Azure
Synapse Analytics
retails-synapse-workspace
Search
rajasuhashkesari@gmail.com
DEFAULT DIRECTORY
Synapse live
Validate all
Publish all 1
Creating External Tab...
Creating View gold o...
Intro OPENROWSET ...
SQL - Scripting usin...
Other users in your workspace may have access to modify this item. Do not use this item unless you trust all users who may have access to the workspace.
Run
Undo
Publish
Query plan
Connect to
Built-in
Use database
olist
1 USE olist
2 GO
3 SELECT TOP 100 * FROM gold.olist_orders_list;
4
5
6 --Total number of products which having atleast one order.
7 SELECT COUNT(DISTINCT(product_id)) AS Total_No_Products_Atleast_One_Order FROM gold.olist_orders_list;
8 GO
9
10 --Total number of quantity of products successfully deliveried.
11 SELECT COUNT((order_item_id)) AS "Number of items Deliveried" FROM gold.olist_orders_list WHERE order_status = 'Delivered';
12 GO
Results
Messages
View
Table
Chart
Export results
Search
product_categ... seller_id product_id order_id customer_id order_status order_purchas... order_approve... order_delivere... order_delivere... order_estimate... actual_delive
automotivo 6a8b085f816a1... d899e614656fe... e659b8223be6f... 9ac71dee8847f... delivered 2017-11-18 2017-11-18 11:... 2017-11-20 14:... 2017-12-08 2017-12-18 20
moveis_decora... 82e0a475a88cc... 035a4019c8009... aa2e81559d88c... 0d49e6cf6c604... delivered 2017-03-27 2017-03-27 14:... 2017-03-30 16:... 2017-05-05 2017-05-02 39
esporte_lazer cf8ab1616079e... b4436da747c3... dd6d0f11a9c3d... e6c386cfd321a... delivered 2018-01-11 2018-01-13 19:... 2018-02-08 21:... 2018-02-21 2018-03-01 41
00:00:10 Query executed successfully.


retails-synapse-workspace - Mi
retails-synapse-workspace - Az
New Tab
web.azuresynapse.net/en/authoring/analyze/sqlscripts/SQL%20-%20Scripting%20using%20EXTB%20olist_orders_list?workspace=%2Fsubscriptions%2F394a429d-397d-4207-af88-ca3b078d5e5...
Microsoft Azure
Synapse Analytics
retails-synapse-workspace
Search
rajasuhashkesari@gmail.com
DEFAULT DIRECTORY
Synapse live
Validate all
Publish all 1
Creating External Tab...
Creating View gold o...
Intro OPENROWSET ...
SQL - Scripting usin...
Other users in your workspace may have access to modify this item. Do not use this item unless you trust all users who may have access to the workspace.
Run
Undo
Publish
Query plan
Connect to
Built-in
Use database
olist
1 USE olist
2 GO
3 SELECT TOP 100 * FROM gold.olist_orders_list;
4
5
6 --Total number of products which having atleast one order.
7 SELECT COUNT(DISTINCT(product_id)) AS Total_No_Products_Atleast_One_Order FROM gold.olist_orders_list;
8 GO
9
10 --Total number of quantity of products successfully deliveried.
11 SELECT COUNT((order_item_id)) AS "Number of items Deliveried" FROM gold.olist_orders_list WHERE order_status = 'Delivered';
12 GO
Results
Messages
View
Table
Chart
Export results
Search
Total_No_Products_Atleast_One_Order
32951
00:00:01 Query executed successfully.



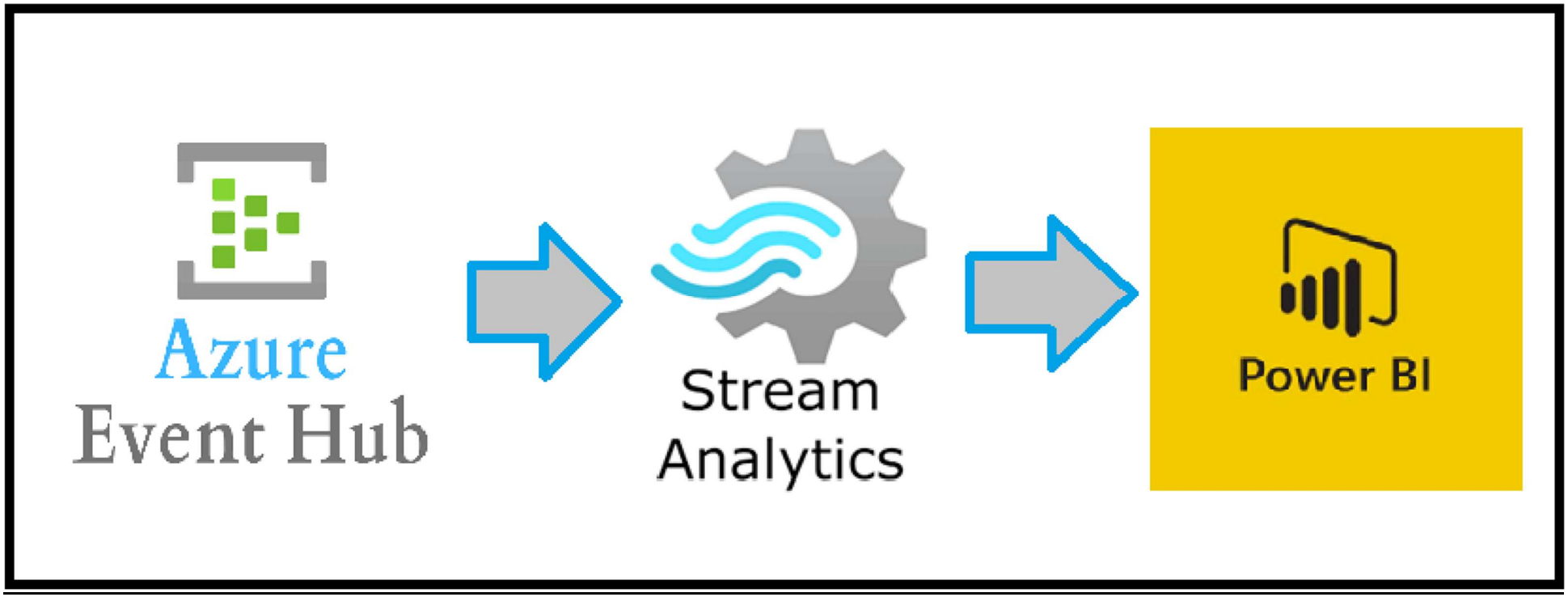
You see the outputs in the below of the picture.

Solution for Implementing

Stream Processing

in Azure Data Engineering

Dataflow and Design :-



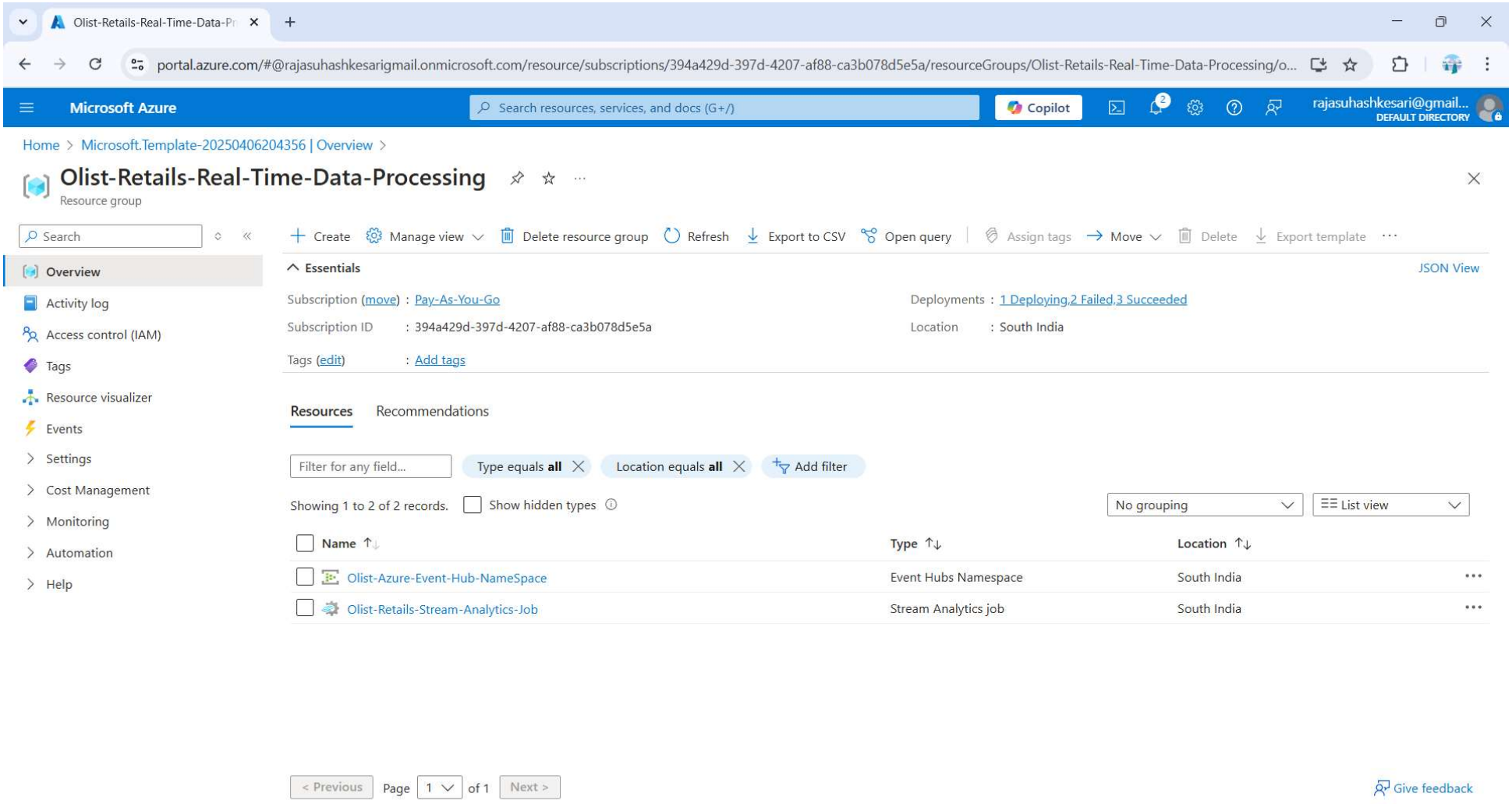
To implement Stream Processing we need 3 services.

1. **Azure Event Hub**
2. **Azure Stream Analytics**
3. **PowerBI**

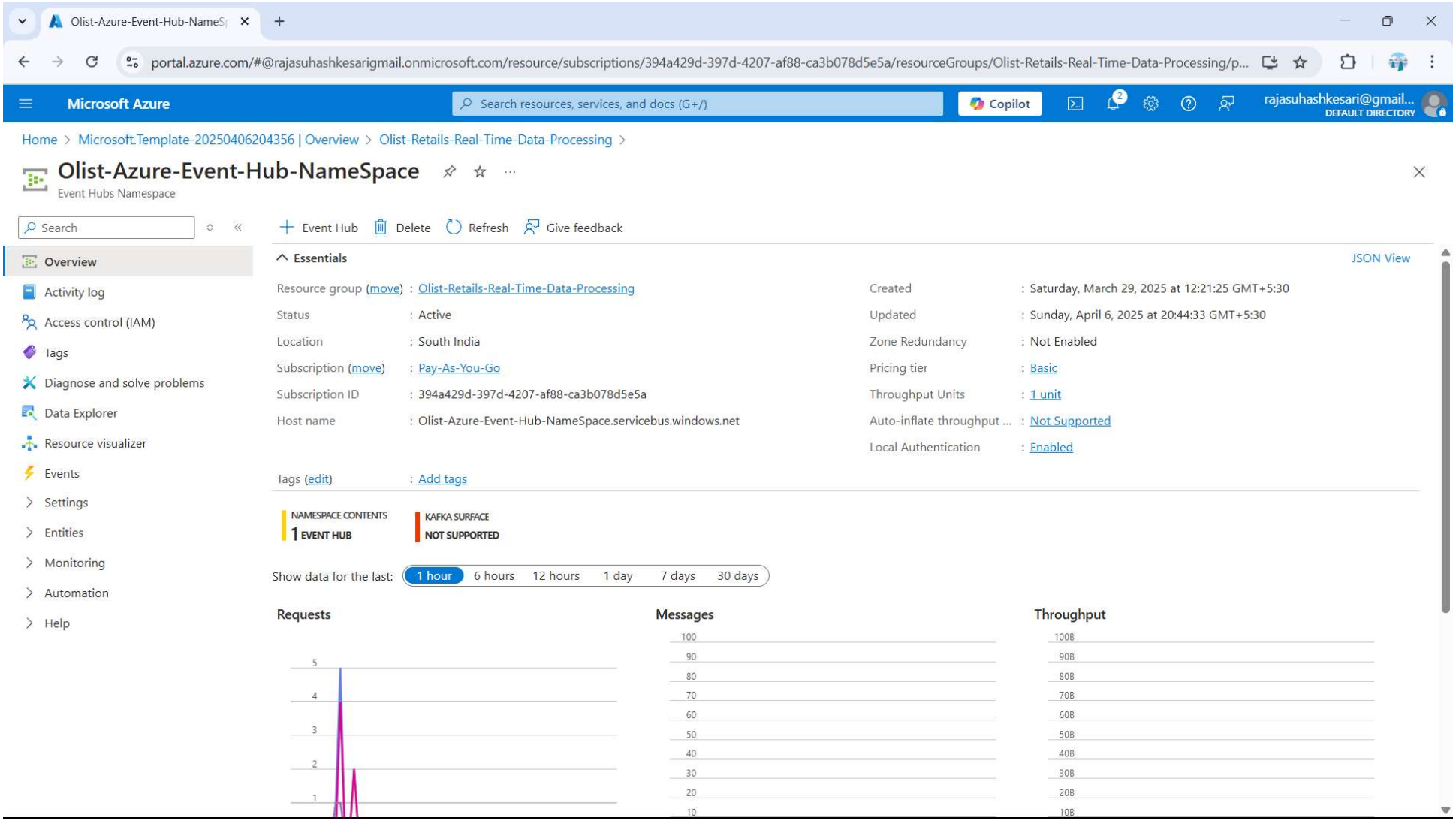
Azure Event Hub acts as the data ingestion service, capturing large volumes of streaming data from various sources. This data is then passed to Azure Stream Analytics, which performs real-time analytics and transformations using SQL-like queries. Finally, the processed output is sent directly to Power BI, where it is visualized in interactive dashboards that refresh in near real-time, providing actionable insights on streaming data. This setup is ideal for monitoring scenarios like IoT telemetry, e-commerce analytics, or social media trends.

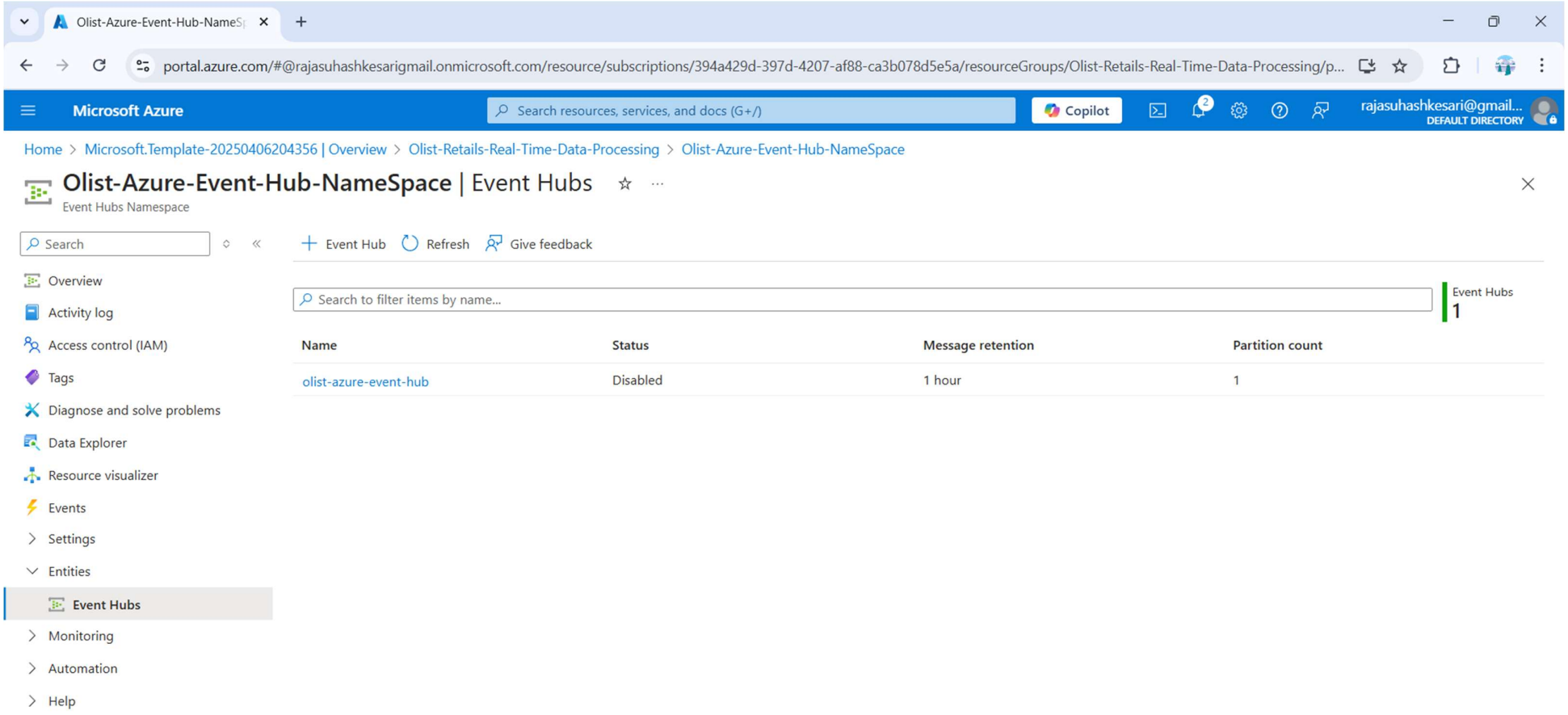
1. Azure Event Hub :-

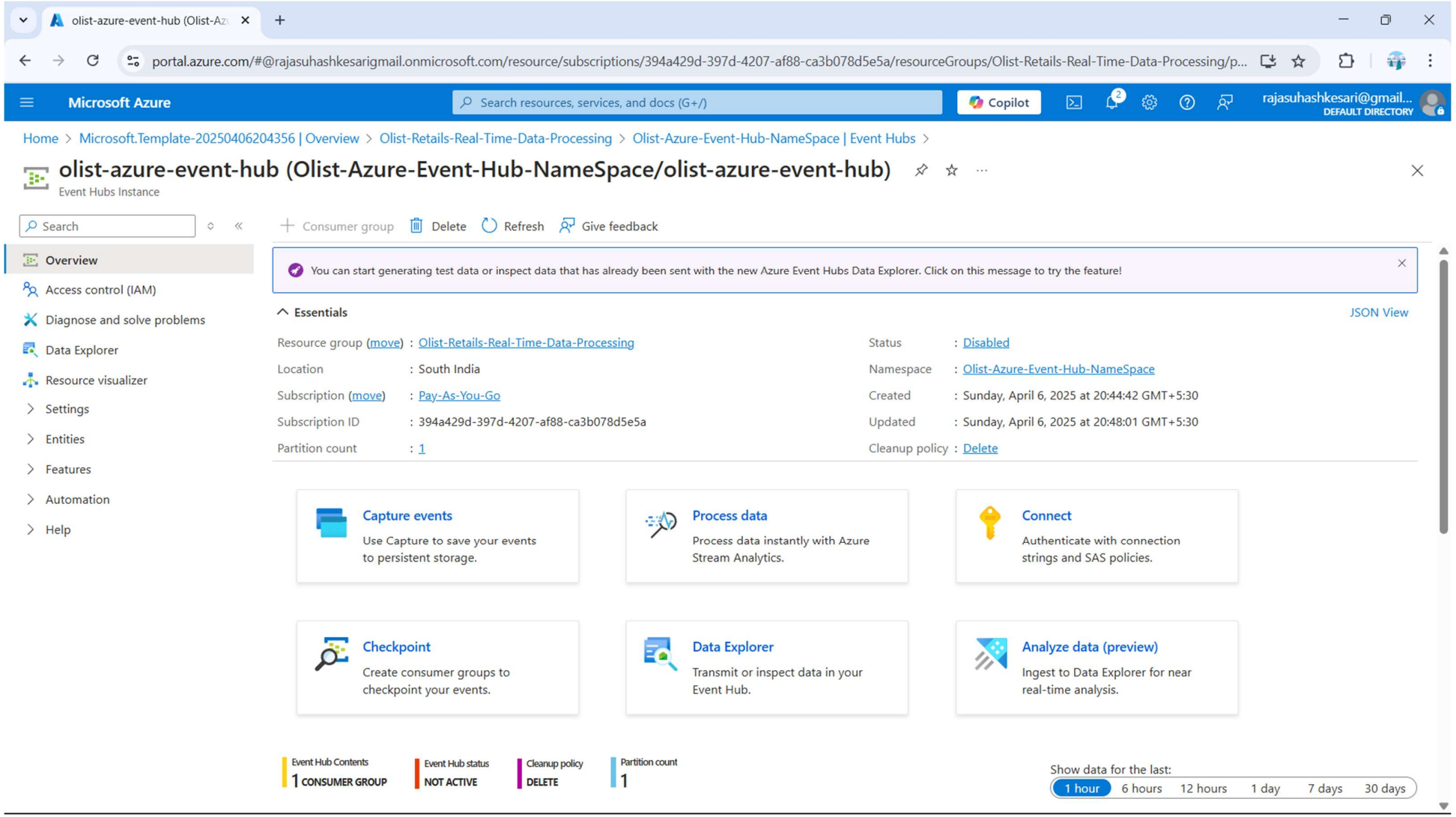
For Stream processing I have created separate resource group name of that is **Olist-Retails-Real-Time-Data-Processing**



Step - 1 :- I have created **‘Azure Event Hub’** with name **‘Olist-Azure-Event-Hub-NameSpace’.**




Olist-Azure-Event-Hub-NameS
portal.azure.com/#@rajasuhashkesarigmail.onmicrosoft.com/resource/subscriptions/394a429d-397d-4207-af88-ca3b078d5e5a/resourceGroups/Olist-Retails-Real-Time-Data-Processing/p...
Microsoft Azure
Search resources, services, and docs (G+/)
Copilot
rajasuhashkesari@gmail...
DEFAULT DIRECTORY
Home > Microsoft.Template-20250406204356 | Overview > Olist-Retails-Real-Time-Data-Processing > Olist-Azure-Event-Hub-NameSpace
Olist-Azure-Event-Hub-NameSpace | Event Hubs
Event Hubs Namespace
Search
Overview
Activity log
Access control (IAM)
Tags
Diagnose and solve problems
Data Explorer
Resource visualizer
Events
Settings
Entities
Event Hubs
Monitoring
Automation
Help
Event Hub
Refresh
Give feedback
Search to filter items by name...
Event Hubs
1
Name
Status
Message retention
Partition count
olist-azure-event-hub
Disabled
1 hour
1

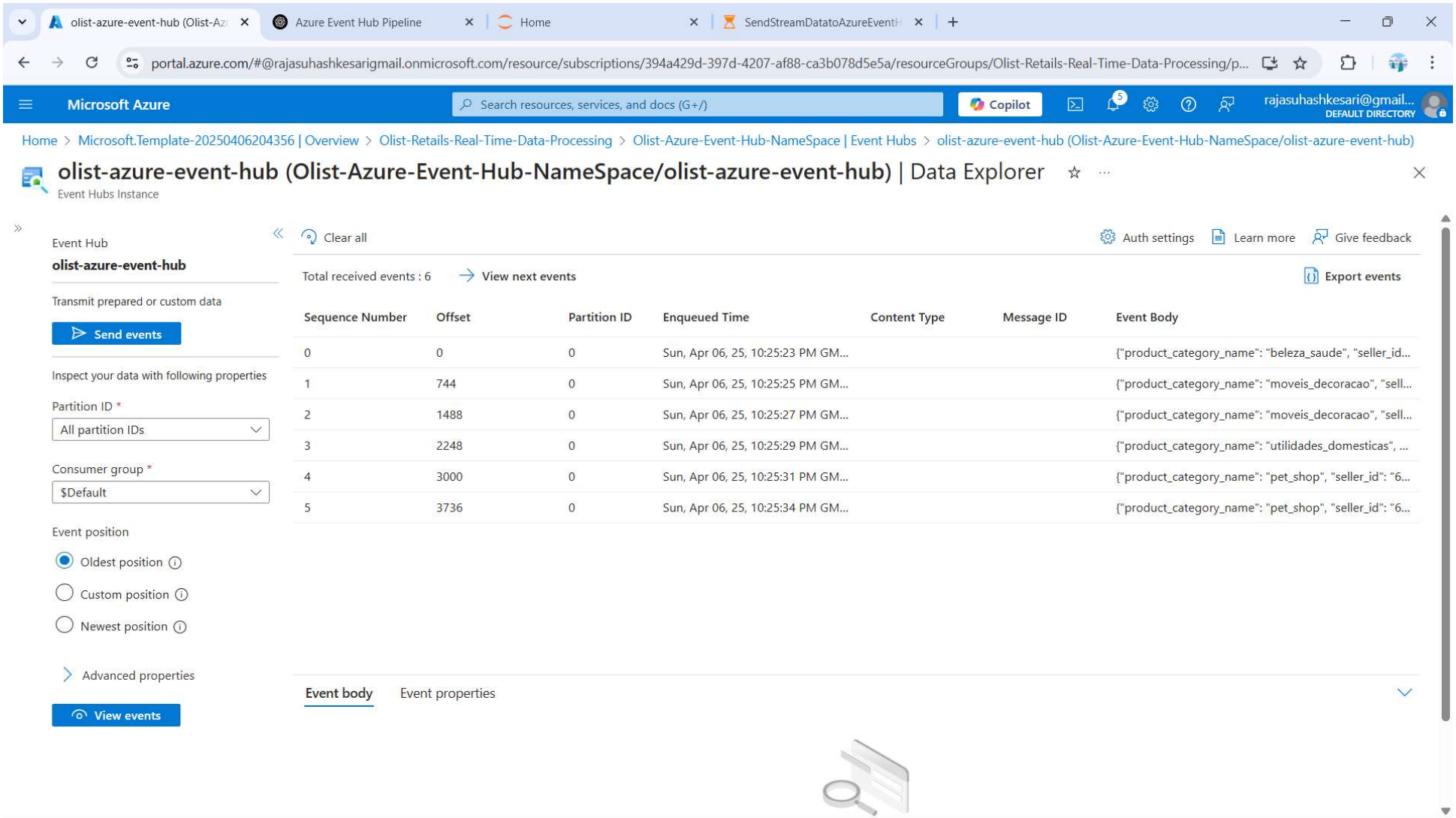

olist-azure-event-hub (Olist-Azu
portal.azure.com/#@rajasuhashkesarigmail.onmicrosoft.com/resource/subscriptions/394a429d-397d-4207-af88-ca3b078d5e5a/resourceGroups/Olist-Retails-Real-Time-Data-Processing/p...
Microsoft Azure
Search resources, services, and docs (G+/)
Copilot
rajasuhashkesari@gmail...
DEFAULT DIRECTORY
Home > Microsoft.Template-20250406204356 | Overview > Olist-Retails-Real-Time-Data-Processing > Olist-Azure-Event-Hub-NameSpace | Event Hubs >
olist-azure-event-hub (Olist-Azure-Event-Hub-NameSpace/olist-azure-event-hub)
Event Hubs Instance
Search
Overview
Access control (IAM)
Diagnose and solve problems
Data Explorer
Resource visualizer
Settings
Entities
Features
Automation
Help
Consumer group
Delete
Refresh
Give feedback
You can start generating test data or inspect data that has already been sent with the new Azure Event Hubs Data Explorer. Click on this message to try the feature!
Essentials
JSON View
Resource group (move) : Olist-Retails-Real-Time-Data-Processing
Location : South India
Subscription (move) : Pay-As-You-Go
Subscription ID : 394a429d-397d-4207-af88-ca3b078d5e5a
Partition count : 1
Status : Disabled
Namespace : Olist-Azure-Event-Hub-NameSpace
Created : Sunday, April 6, 2025 at 20:44:42 GMT+5:30
Updated : Sunday, April 6, 2025 at 20:48:01 GMT+5:30
Cleanup policy : Delete
Capture events
Use Capture to save your events to persistent storage.
Process data
Process data instantly with Azure Stream Analytics.
Connect
Authenticate with connection strings and SAS policies.
Checkpoint
Create consumer groups to checkpoint your events.
Data Explorer
Transmit or inspect data in your Event Hub.
Analyze data (preview)
Ingest to Data Explorer for near real-time analysis.
Event Hub Contents
1 CONSUMER GROUP
Event Hub status
NOT ACTIVE
Cleanup policy
DELETE
Partition count
1
Show data for the last:
1 hour
6 hours
12 hours
1 day
7 days
30 days

This is the python code which is going send the live data to **azure event hub.**

```
[1]: import pandas as pd
     import json
     from azure.eventhub import EventHubProducerClient, EventData
     import random
     import time as t
     import numpy as np

[2]: order_ds = pd.read_csv("real_time_orders_dataset.csv")

[5]: con_string = "Endpoint=sb://olist-azure-event-hub-namespace.servicebus.windows.net/;SharedAccessKeyName=Manage-policy;SharedA
     event_hub_name = "olist-azure-event-hub"

[*]: row_finished=0
     while row_finished <= order_ds.shape[0]:
         random_number = random.randint(1, 10)
         for i in range(random_number):
             order = {key: (int(value) if isinstance(value, (np.int64, np.float64)) else value) for key, value in order_ds.iloc[i+
             print(order)
             message = json.dumps(order)  # Convert dictionary to JSON string
             producer = EventHubProducerClient.from_connection_string(con_string, eventhub_name=event_hub_name)
             event_data = EventData(message)
             producer.send_batch([event_data])
         row_finished = row_finished + random_number
         t.sleep(0.1)

     {'product_category_name': 'beleza_saude', 'seller_id': '2a5b78b41cd05baeac8df54c6606b92c', 'product_id': '7fbf3cba00dcf6da
     3f7c0fe6a162b4c1', 'order_id': '908e767e28908dc5ec61fe0ef8ef03f9', 'customer_id': '64d181ef52b1829800c598942459061b', 'cus
     tomer_unique_id': 'd903cb2df3fc0df3e432bc359d8a6bf2', 'customer_zip_code_prefix': 37903, 'customer_city': 'passos', 'custo
```

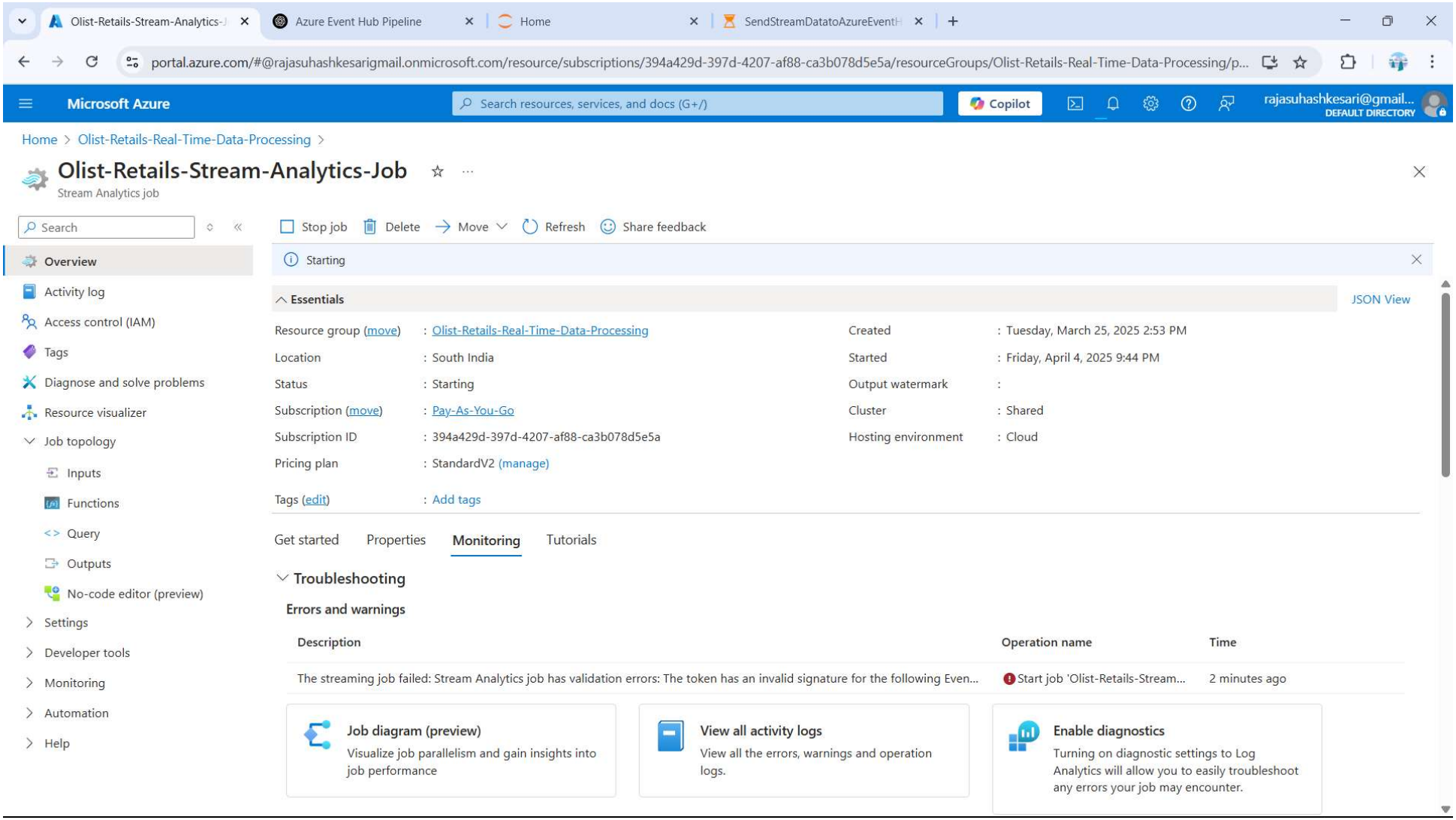
the python code is running the events are sending to azure event hub

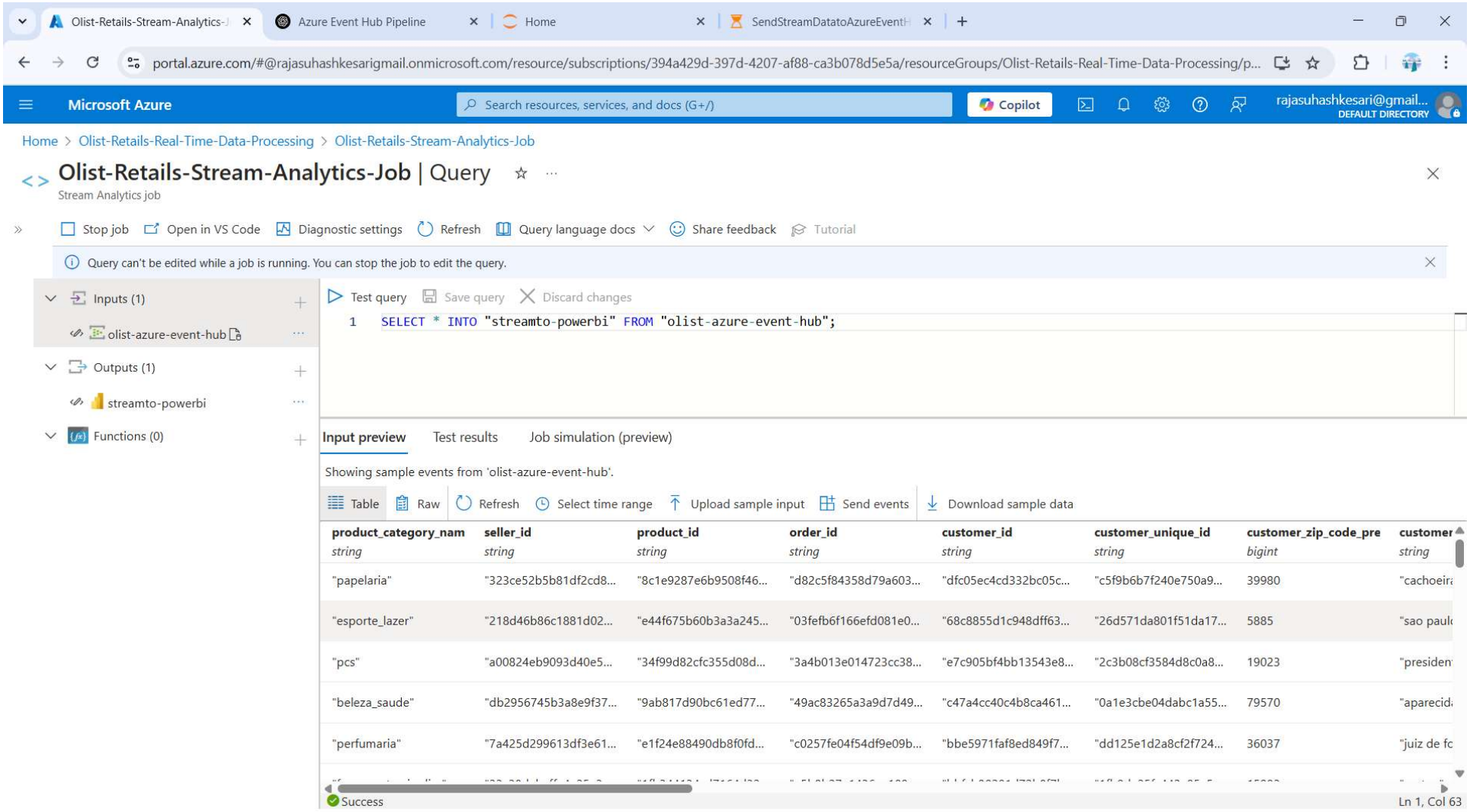
At azure event hub side we are receving events bewlo are the events recevied .



Azure Stream Analytics:-

This is the my **“azure stream analytics job”** running

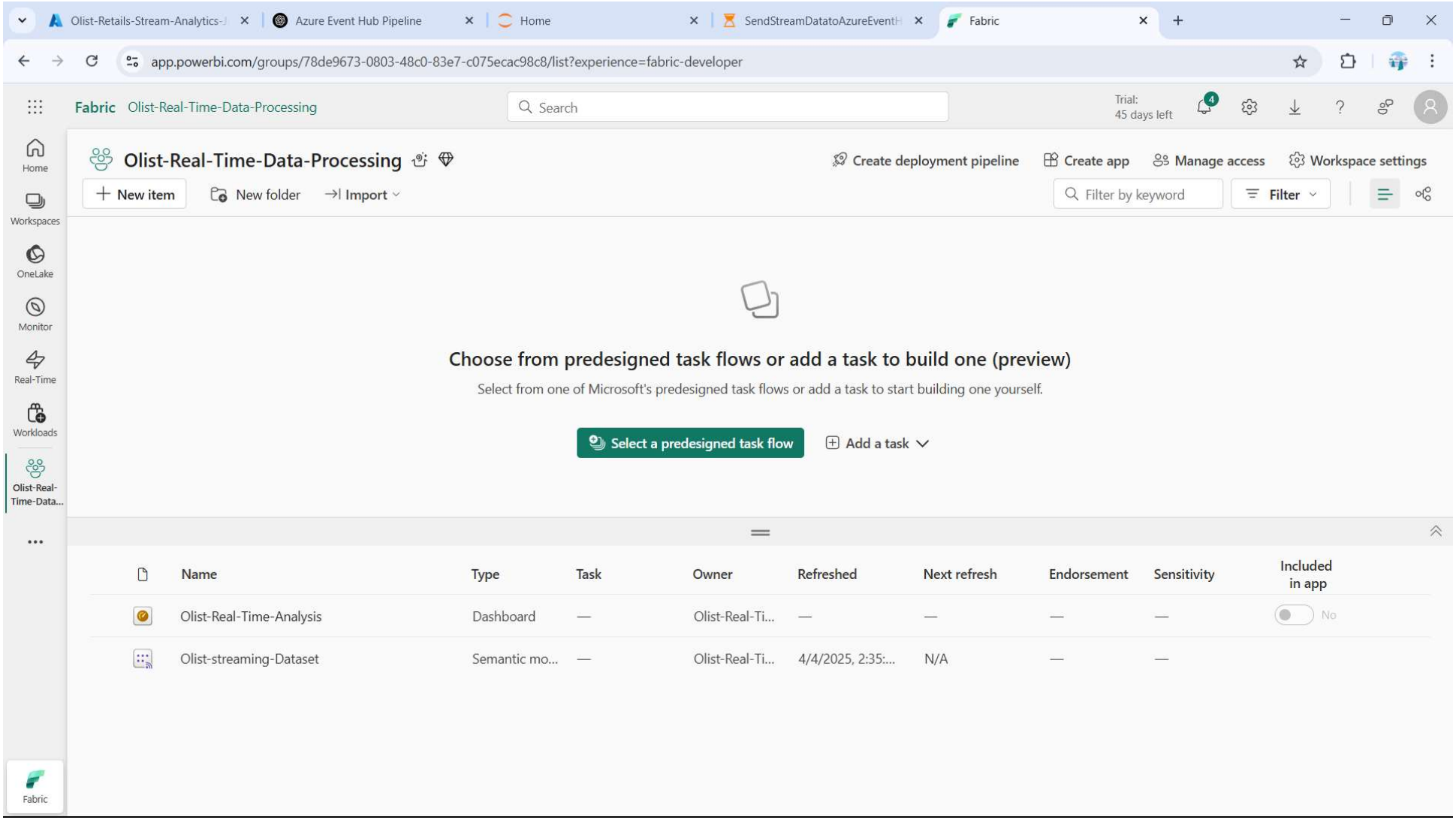




I have configured input and outputs stream analytics job and it acts as mediator between azure event hub and powerbi.

It takes events from the event hub and send it the powerBI . by using sql code we can transform the data we can perform aggregation but I haove not done it. Iam just directly seding realtime data to powerbi.

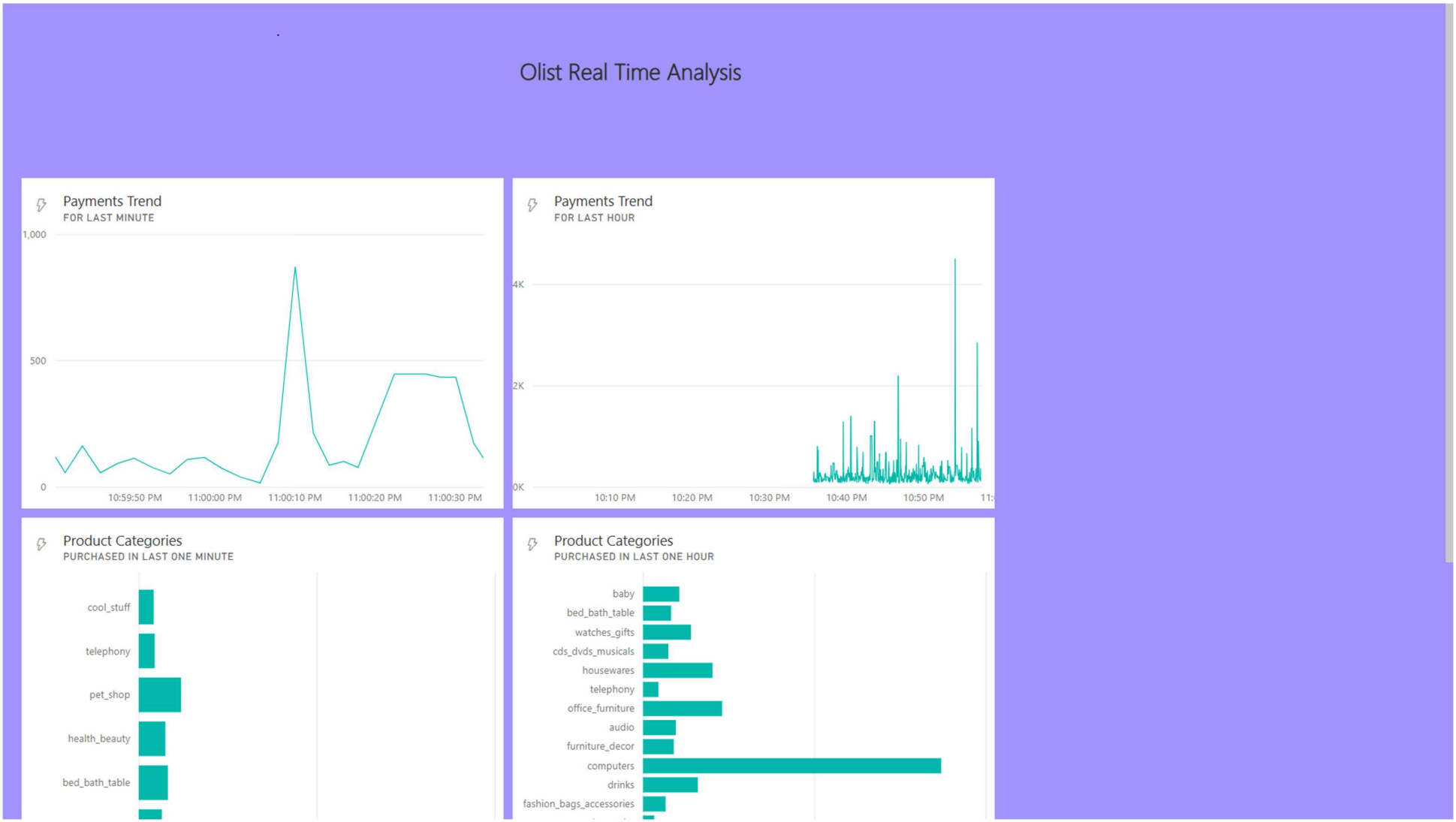
PowerBI :-



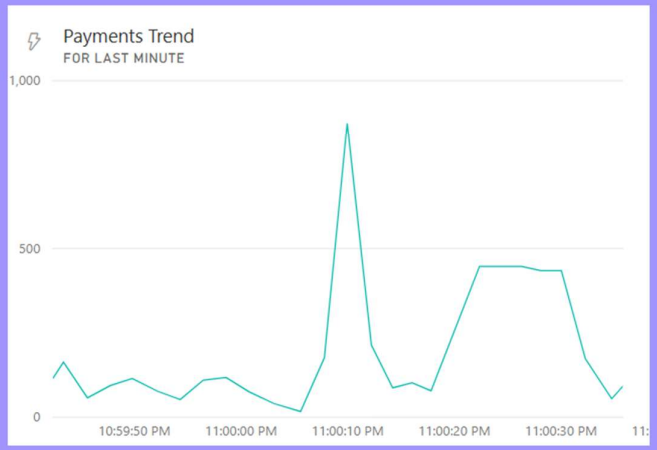
In the above Olist-streaming-Dataset is the semantic model which is receiving the Stream data.

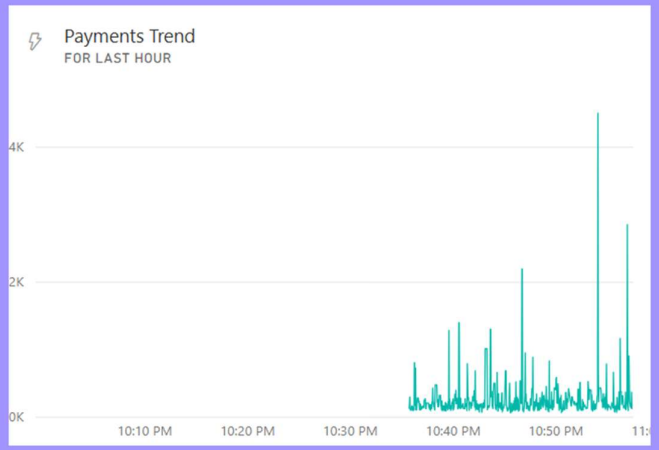
From the I have created a PowerBI Dashboard which ‘Olist-Real-Time-Analysis’. In that we see live dashboard.

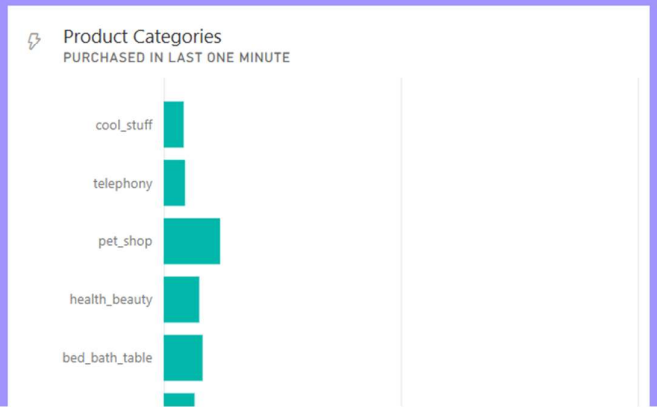
This is the live dashboard in which some tile will show real time analysis for last 1 minute and last 1 hour.

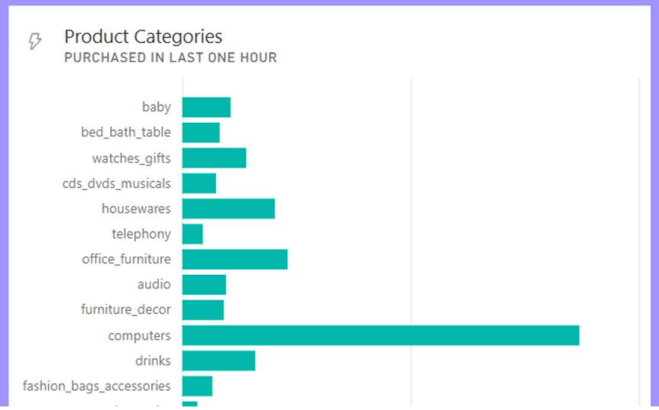


Olist Real Time Analysis

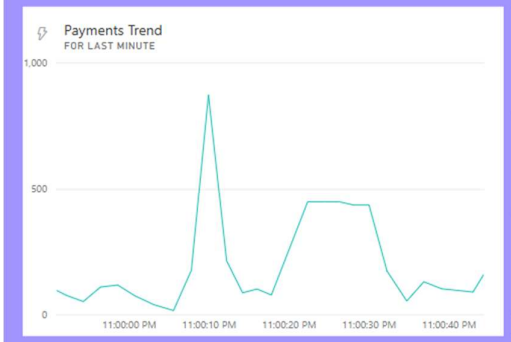


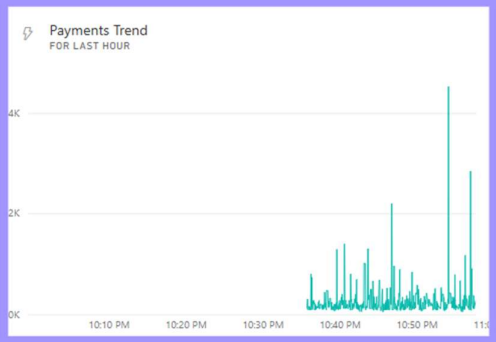


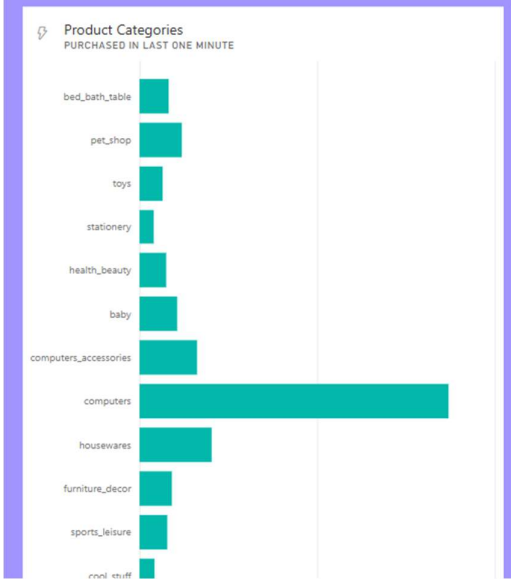


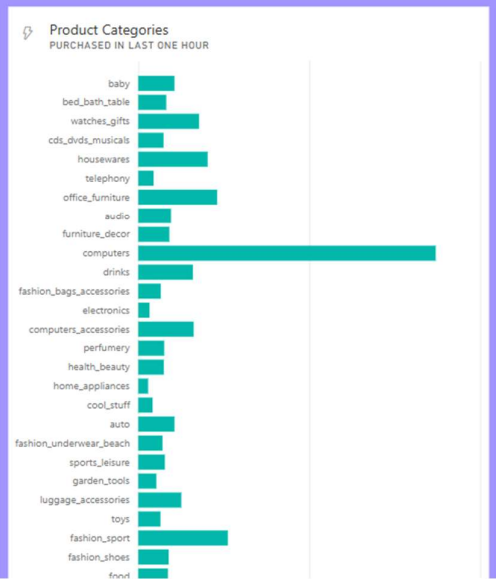


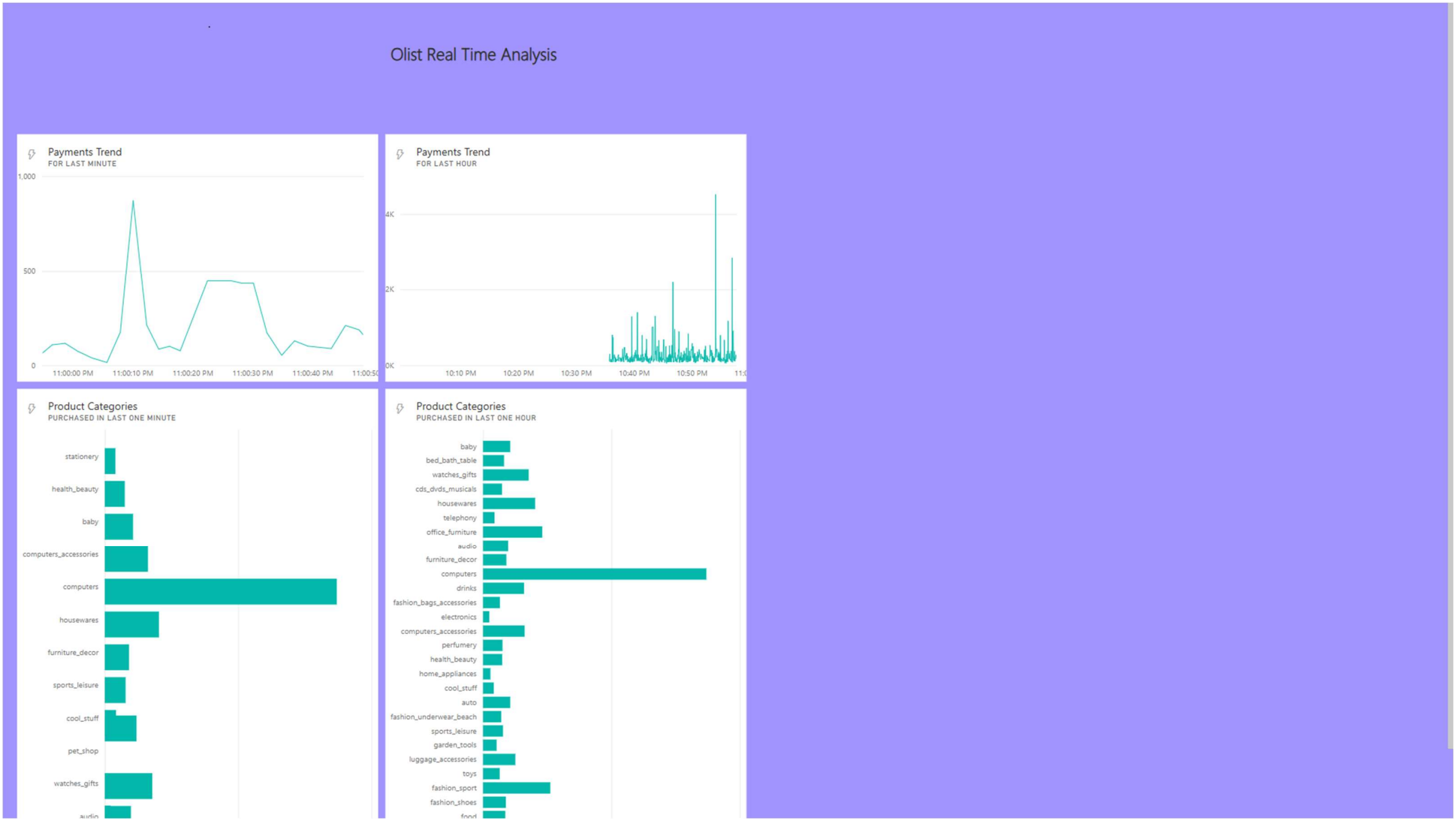
Olist Real Time Analysis











We can observe the data in the above tale was refreshing .

The End

Thank You