

Microsoft Azure Data Engineering Training Course

Capstone Project

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

Objectives:-

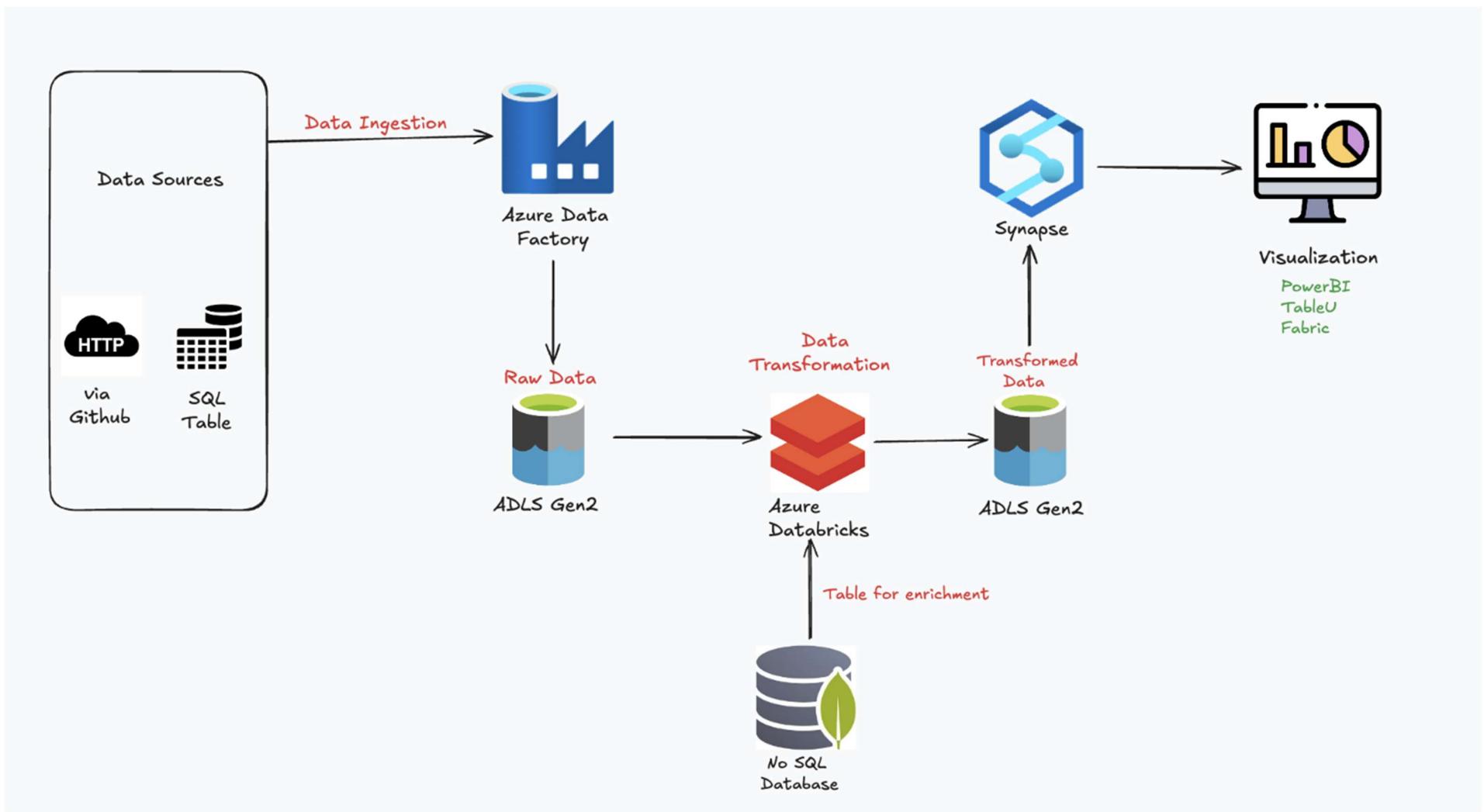
- 1. Implement Batch Processing in Azure Data Engineering Services for Online Retails Dataset.**
- 2. Implement Stream Processing in Azure Data Engineering Services for Online Retails Streaming Dataset.**

Solution for Implementing

Batch Processing

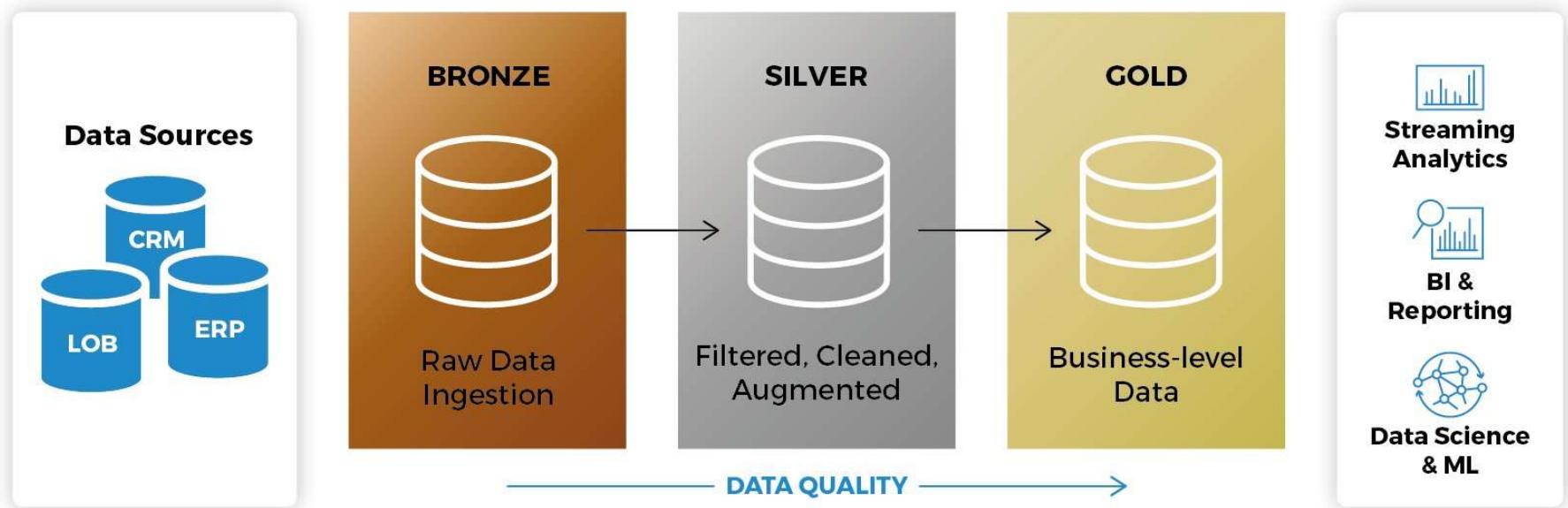
in Azure Data Engineering

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 = 00143d0f86d6fb9f9b38ab440ac16f5 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.

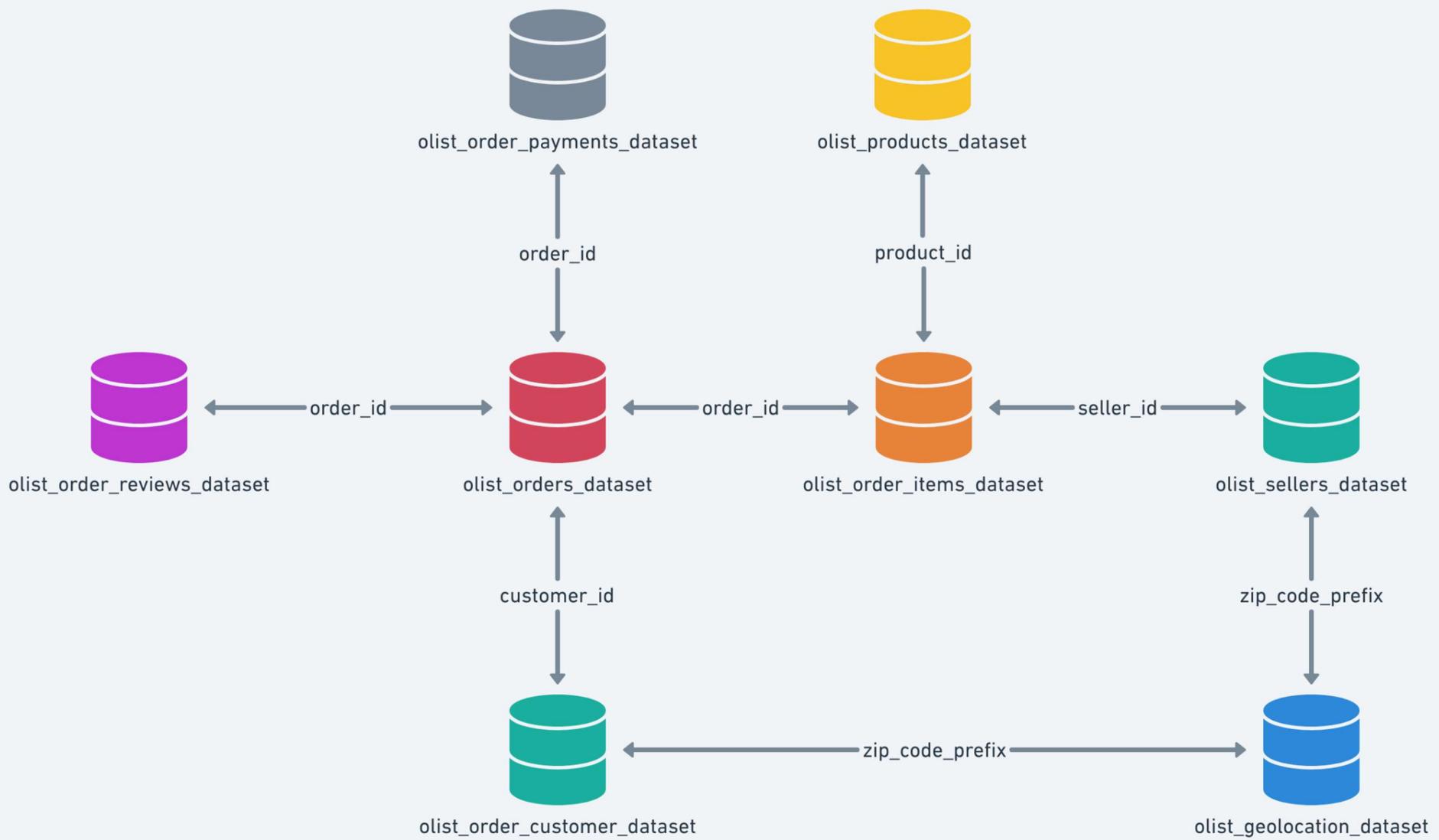
8.Sellers Dataset

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:-



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’

The screenshot shows the Microsoft Azure portal interface. The top navigation bar has two tabs: 'Data-Driven-Decision-Making' and 'Copilot'. The address bar shows the URL: portal.azure.com/#@rajasuhashkesarigmail.onmicrosoft.com/resource/subscriptions/394a429d-397d-4207-af88-ca3b078d5e5a/resourceGroups/Data-Driven-Decision-Making-for-Enhanced-Online-Retail-Experience-and-Increased-Sales. The top right corner shows the user's email: rajasuhashkesari@gmail.com and the directory: DEFAULT DIRECTORY (RAJASUHA...).

The main page title is 'Data-Driven-Decision-Making-for-Enhanced-Online-Retail-Experience-and-Increased-Sales' under the 'Resource group' category. The left sidebar contains navigation links: Home, Overview, Essentials, Activity log, Access control (IAM), Tags, Resource visualizer, Events, Settings, Cost Management, Monitoring, Automation, and Help.

The 'Essentials' section is selected. It includes a search bar, a 'Create' button, and various management options like 'Manage view', 'Delete resource group', 'Refresh', 'Export to CSV', 'Open query', 'Assign tags', 'Move', 'Delete', and 'Export template'. There is also a 'JSON View' link.

The 'Resources' tab is active, showing 'Recommendations (1)'. Below it is a filter bar with 'Filter for any field...', 'Type equals all', 'Location equals all', and an 'Add filter' button. A note says 'Showing 1 to 4 of 4 records.' with a checkbox for 'Show hidden types'.

The main content area displays a table of resources:

Name ↑↓	Type ↑↓	Location ↑↓	...
online-retails-azure-databricks-premium	Azure Databricks Service	Central India	...
Retails-Data-Factory	Data factory (V2)	Central India	...
retails-synapse-workspace	Synapse workspace	Central India	...
retailsadlsgen2	Storage account	Central India	...

At the bottom, there are navigation buttons: '< Previous', 'Page 1 of 1', 'Next >', and a 'Give feedback' link. The bottom right corner shows system status: ENG IN, 15:05, 05-04-2025, and a battery icon.

Step 2:-Created Storage Account with `retailsadlsgen2` name.

I created Data Lake Storage Gen2.

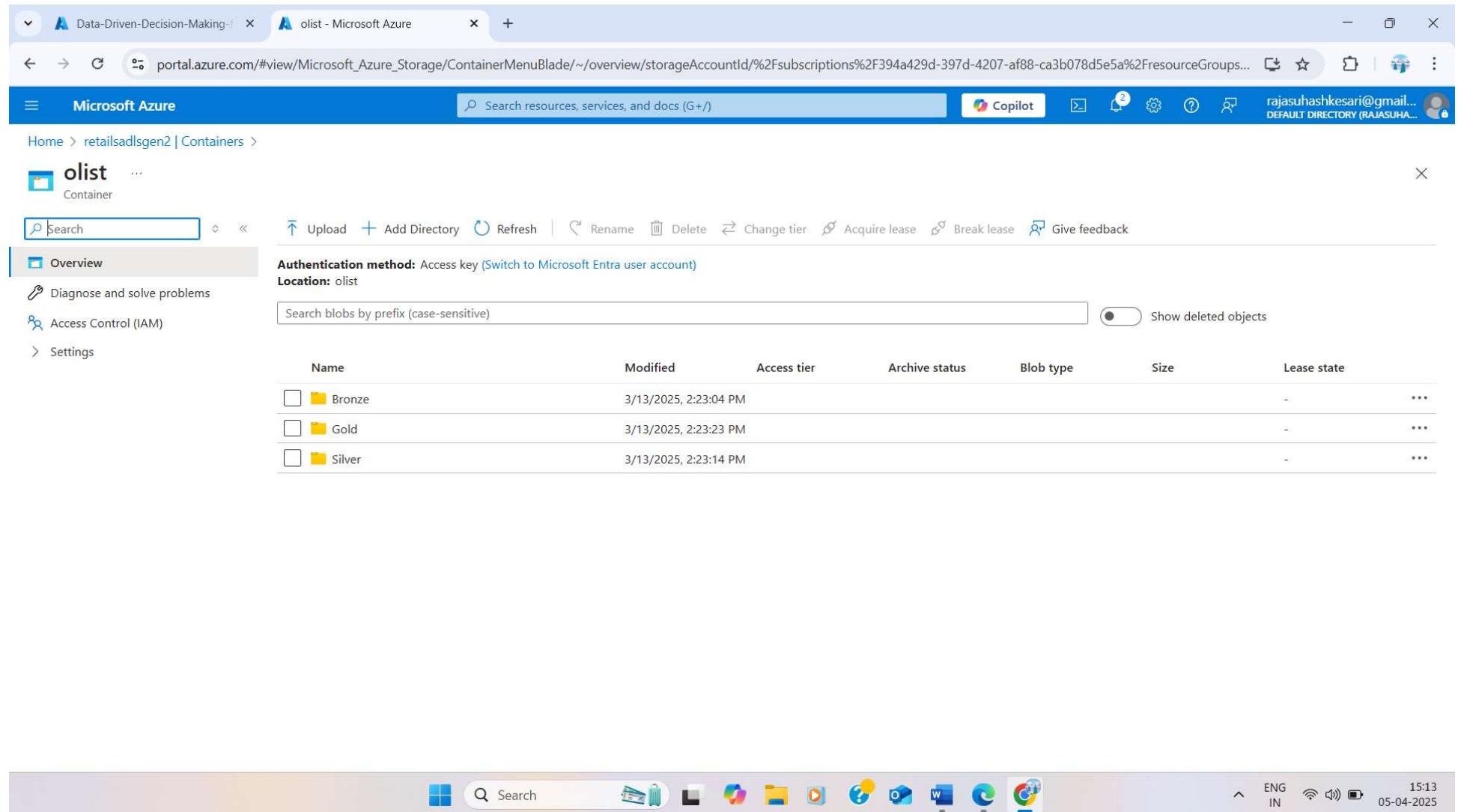
The screenshot shows the Microsoft Azure portal interface for a storage account named 'retailsadlsgen2'. The left sidebar contains navigation links for Overview, Activity log, Tags, Diagnose and solve problems, Access Control (IAM), Data migration, Events, Storage browser, Partner solutions, Resource visualizer, Data storage (Containers, File shares, Queues, Tables), Security + networking, Data management, and Settings. The main content area displays the 'Essentials' section with details like Resource group, Location (centralindia), Subscription ID, Disk state, and Tags. It also shows the 'Properties' tab with sections for Data Lake Storage (Hierarchical namespace: Enabled, Default access tier: Hot, Blob anonymous access: Disabled, Blob soft delete: Disabled, Container soft delete: Disabled, Versioning: Disabled, Change feed: Disabled, NFS v3: Disabled, SFTP: Disabled) and Security (Require secure transfer for REST API operations: Enabled, Storage account key access: Enabled, Minimum TLS version: Version 1.2, Infrastructure encryption: Disabled). The Networking section indicates Allow access from All networks and Private endpoint connections 0. The top navigation bar includes tabs for Data-Driven-Decision-Making-1 and retailads1gen2 - Microsoft Azure, and a search bar for 'Search resources, services, and docs (G+/-)'. The bottom taskbar shows various application icons and system status indicators.

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

The screenshot shows the Microsoft Azure portal interface for a storage account named 'retailsadlsgen2'. The left sidebar navigation bar is visible, with 'Containers' selected under the 'Data storage' category. The main content area displays a list of containers. The container 'olist' is selected, indicated by a checked checkbox next to its name. Other containers listed are 'retailsfilesystem' and another unnamed container. The table columns include Name, Last modified, Anonymous access level, and Lease state. The 'Anonymous access level' for 'olist' is set to 'Private'.

Name	Last modified	Anonymous access level	Lease state
olist	3/13/2025, 2:22:43 PM	Private	Available
retailsfilesystem	3/11/2025, 1:05:50 PM	Private	Available

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



The screenshot shows the Microsoft Azure Storage Container Overview page for the 'olist' container. The left sidebar includes links for Overview, Diagnose and solve problems, Access Control (IAM), and Settings. The main area displays a table of blobs with the following data:

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
Bronze	3/13/2025, 2:23:04 PM				-	...
Gold	3/13/2025, 2:23:23 PM				-	...
Silver	3/13/2025, 2:23:14 PM				-	...

At the bottom of the page, there is a toolbar with icons for Upload, Add Directory, Refresh, Rename, Delete, Change tier, Acquire lease, Break lease, and Give feedback. The status bar at the bottom right shows the date (05-04-2025), time (15:13), language (ENG IN), and battery level.

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

The screenshot shows the Microsoft Azure portal interface. The top navigation bar includes tabs for 'Data-Driven-Decision-Making...' and 'Retails-Data-Factory - Microsoft'. The address bar displays the URL: portal.azure.com/#@rajasuhashkesarigmail.onmicrosoft.com/resource/subscriptions/394a429d-397d-4207-af88-ca3b078d5e5a/resourceGroups/Data-Driven-Decision-Making-for-Enhanc.... The main header bar has a search bar, Copilot, and other account-related icons. The left sidebar shows the 'Retails-Data-Factory' Data factory (V2) with a list of options: Overview (selected), Activity log, Access control (IAM), Tags, Diagnose and solve problems, Resource visualizer, Settings, Getting started, Monitoring, Automation, and Help. The 'Overview' section displays the following details:

Essentials	
Resource group (move)	: Data-Driven-Decision-Making-for-Enhanced-Online-Retail-Experienc...
Status	: Succeeded
Location	: Central India
Subscription (move)	: Pay-As-You-Go
Subscription ID	: 394a429d-397d-4207-af88-ca3b078d5e5a

The 'Type' is listed as 'Data factory (V2)' and 'Getting started' is linked to 'Quick start'. Below this information is a large blue icon of a factory building. The text 'Azure Data Factory Studio' is centered above a blue button labeled 'Launch studio'. At the bottom of the page, there are four cards: 'Quick Starts' (cloud icon), 'Tutorials' (book icon), 'Template Gallery' (document icon), and 'Training Modules' (certificate icon). The bottom navigation bar includes a 'Monitoring' tab, a search bar, and various system icons like battery level and signal strength. The status bar at the bottom right shows the time as 15:16 and the date as 05-04-2025.

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

The screenshot shows the Microsoft Azure Data Factory studio interface. At the top, there are three browser tabs: "Data-Driven-Decision-Making", "Retails-Data-Factory - Microsoft Edge", and "Retails-Data-Factory - Azure Data Factory". The URL in the address bar is adf.azure.com/en/home?factory=%2Fsubscriptions%2F394a429d-397d-4207-af88-ca3b078d5e5a%2FresourceGroups%2FDATA-Driven-Decision-Making-for-Enhanced-Online-Retail-Experience-%2Fdatafactories%2FRetails-Data-Factory. The top navigation bar includes "Microsoft Azure", "Data Factory", "Retails-Data-Factory", a search bar, and user information for "rajasuhashkesari@gmail.com" in the "DEFAULT DIRECTORY".

The main content area displays the "Retails-Data-Factory" dashboard. It features a large central graphic illustrating data flow from various sources (represented by blue cylinders) into a central processing area (represented by a factory building), which then outputs data to a destination (represented by a blue cylinder). Below this graphic are four cards:

- Ingest**: Copy data at scale once or on a schedule.
- Orchestrate**: Code-free data pipelines.
- Transform data**: Transform your data using data flows.
- Configure SSIS**: Manage & run your SSIS packages in the cloud.

Below these cards, under the heading "Recent resources", is a table listing five recent items:

Name	Type	Last opened by you
PL_source_to_bronze	Pipeline	03/26/2025
pl_source_to_bronze	Pipeline	03/13/2025
DS MySql_order_payments	Dataset	03/13/2025
DS_olist_bronze_csv_file	Dataset	03/13/2025
DS_source_metadata_json	Dataset	03/13/2025

The bottom of the screen shows the Windows taskbar with icons for File Explorer, Task View, and other system utilities. The system tray indicates the date as 05-04-2025, the time as 15:21, and language as ENG IN.

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

Edit linked service

HTTP Learn more 

Name * LS_for_Github

Description

Connect via integration runtime *  AutoResolveIntegrationRuntime

Base URL * https://raw.githubusercontent.com/

 Information will be sent to the URL specified. Please ensure you trust the URL entered.

Server certificate validation 
 Enable Disable

Authentication type * 
Anonymous

Auth headers 
+ New

Annotations
+ New

 Test connection

❖Created MySQL linked service and LS_for_SQL_database as its name

Edit linked service

MySQL [Learn more](#)

Name *

Description

Connect via integration runtime * ⓘ

AutoResolveIntegrationRuntime

Server name *

Port

Database name *

User name *

Password Azure Key Vault

Password *

Save Cancel Test connection

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

Edit linked service

 Azure Data Lake Storage Gen2 [Learn more](#)

Name *

LS_AzureDataLakeStorageGen2

Description

Connect via integration runtime * ⓘ

AutoResolveIntegrationRuntime

Authentication type

Account key

Account selection method ⓘ

From Azure subscription Enter manually

URL *

https://retailsadlsgen2.dfs.core.windows.net/

Storage account key

Azure Key Vault

Storage account key *

.....

Test connection ⓘ

To linked service To file path

After Creating we need 4 Datasets so I created them

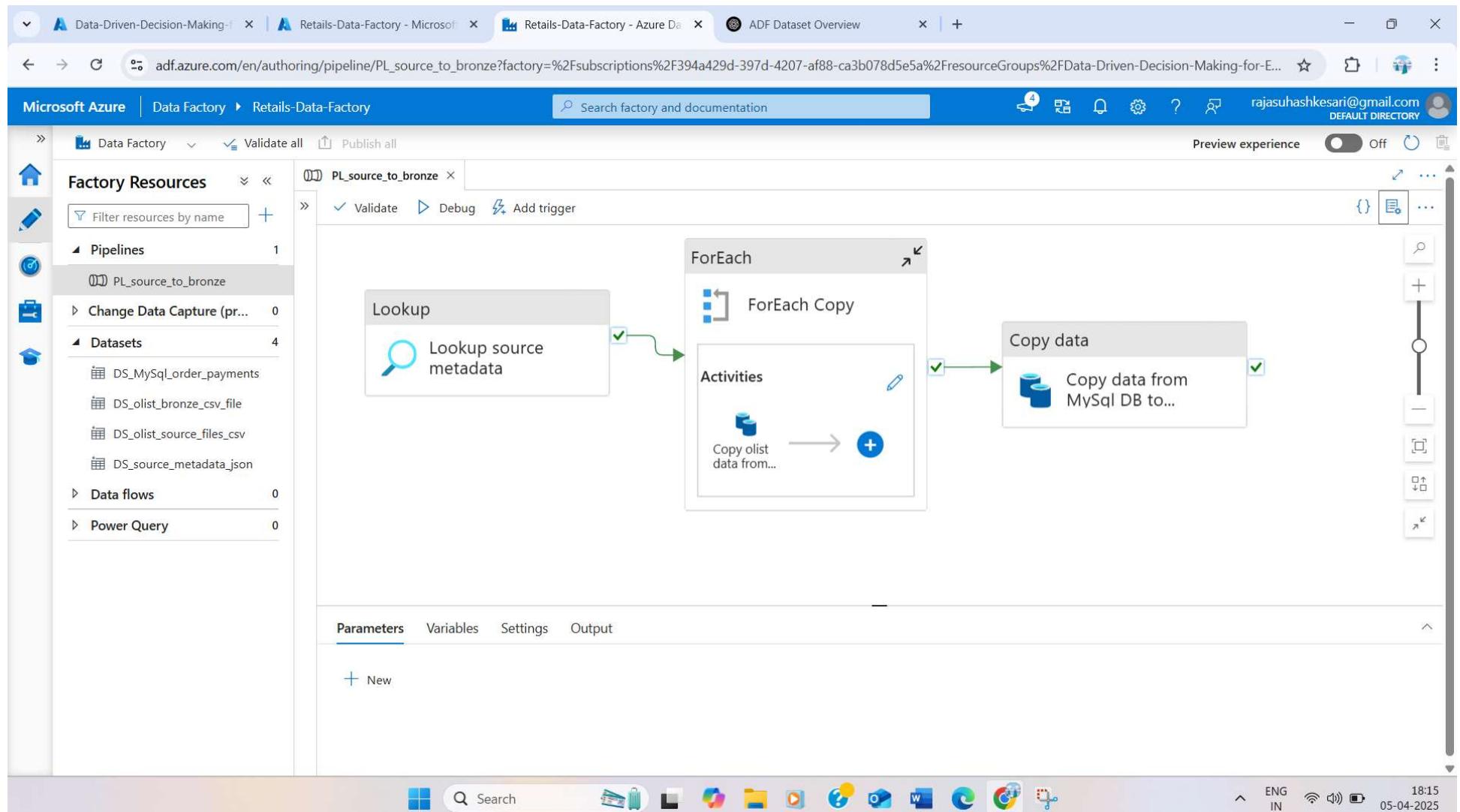
1. **DS MySql_order_payments** :-This dataset likely refers to **order payments data stored in a MySQL database**.
2. **DS_olist_bronze_csv_file**:-This dataset is pointing to the location where the data is going to be loaded after ingestion .
3. **DS_olist_source_files_csv**:- This dataset is refers 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**.

◀ Datasets

4

-  DS MySql_order_payments
-  DS_olist_bronze_csv_file
-  DS_olist_source_files_csv
-  DS_source_metadata_json

--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 pipeline 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 screenshot shows the Microsoft Azure Storage Explorer interface. The left sidebar shows the 'olist' container under 'Bronze'. The main area displays a table of blobs:

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
[..]	3/13/2025, 6:33:15 PM	Hot (Inferred)		Block blob	8.62 MiB	Available
olist_customers_dataset.csv	3/13/2025, 6:33:20 PM	Hot (Inferred)		Block blob	58.44 MiB	Available
olist_geolocation_dataset.csv	3/13/2025, 6:33:31 PM	Hot (Inferred)		Block blob	14.72 MiB	Available
olist_order_items_dataset.csv	3/13/2025, 6:34:26 PM	Hot (Inferred)		Block blob	6.28 MiB	Available
olist_order_payments_dataset.csv	3/13/2025, 6:33:37 PM	Hot (Inferred)		Block blob	13.78 MiB	Available
olist_order_reviews_dataset.csv	3/13/2025, 6:33:48 PM	Hot (Inferred)		Block blob	16.84 MiB	Available
olist_orders_dataset.csv	3/13/2025, 6:33:51 PM	Hot (Inferred)		Block blob	2.27 MiB	Available
olist_products_dataset.csv	3/13/2025, 6:34:01 PM	Hot (Inferred)		Block blob	170.61 KiB	Available

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.

The screenshot shows the Microsoft Azure portal interface. The top navigation bar includes the Microsoft Azure logo, a search bar, and various navigation icons. The main content area displays the details of a Databricks service named 'online-retails-azure-databricks-premium'. The 'Overview' tab is selected, showing the following information:

Essentials	
Status	: Active
Resource group	: Data-Driven-Decision-Making-for-Enhanced-Online-Retail-Experience-and-Increased-Sales
Location	: Central India
Subscription	: Pay-As-You-Go
Subscription ID	: 394a429d-397d-4207-af88-ca3b078d5e5a
Tags (edit)	: Add tags

Managed Resource Group : [online-retails-azure-databricks-premium-mng-r-g](#)
URL : <https://adb-2702396966592863.3.azuredatabricks.net>
Pricing Tier : [Trial \(Premium - 14-Days Free DBUs\) \(Click to change\)](#)

Below the essentials section, there is a large red icon representing three stacked cubes. A blue button labeled 'Launch Workspace' is positioned above a white button labeled 'Upgrade to Premium'.

At the bottom of the page, there are four links: 'Documentation' (with a file icon), 'Getting Started' (with a lightning bolt icon), 'Import Data from File' (with a folder icon), and 'Import Data from Azure Storage' (with a cloud icon). The bottom navigation bar includes the Windows Start button, a search bar, and various system icons. The status bar at the bottom right shows the date (05-04-2025), time (18:46), language (ENG IN), and battery level.

This is my Azure Databricks workspace.

The screenshot shows the Azure Databricks workspace interface. At the top, there are two browser tabs: 'online-retails-azure-databricks-' and 'Home - Databricks'. The URL in the address bar is `adb-2702396966592863.3.azuredatabricks.net/browse?o=2702396966592863`. The page title is 'Home - Databricks'.

The left sidebar contains a navigation menu with the following sections:

- New**
- Workspace** (selected)
- Recents
- Catalog
- Workflows
- Compute
- Marketplace
- SQL
- SQL Editor
- Queries
- Dashboards
- Genie
- Alerts
- Query History
- SQL Warehouses
- Data Engineering
- Job Runs
- Data Ingestion
- Pipelines
- Machine Learning
- Playground

The main content area shows the user profile 'rajasuhashkesari@gmail.com' with a star icon. Below it is a table listing a single notebook:

Name	Type	Owner	Created at
Bronze_to_Silver	Notebook	20EM1A6117 suhash	Mar 23, 2025, 05:20 PM

At the bottom, there is a taskbar with various icons and system status indicators. The system status includes 'ENG IN', a battery icon, signal strength, and the date '05-04-2025'.

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

The screenshot shows the 'Cluster Details - Databricks' page in a browser. The URL is <adb-2702396966592863.3.azuredatabricks.net/compute/clusters/0323-110657-5jw3b0kf?o=2702396966592863>. The left sidebar has 'Compute' selected. The main area shows the configuration for the 'Olist Data Processing Cluster'. The 'Configuration' tab is active. Key settings include:

- Policy:** Unrestricted
- Access mode:** Dedicated (formerly: Single user) - assigned to '20EM1A6117 suhash'
- Performance:** Databricks Runtime Version: 15.4 LTS (includes Apache Spark 3.5.0, Scala 2.12). Includes an option to 'Use Photon Acceleration'.
- Node type:** Standard_F4 - 8 GB Memory, 4 Cores. Includes an option to 'Terminate after 10 minutes of inactivity'.
- Tags:** No custom tags. Includes an 'Automatically added tags' section and an 'Advanced options' section.

Summary:

1 Driver	8 GB Memory, 4 Cores
Runtime	15.4.x-scala2.12
Unity Catalog	Standard_F4
	0.5 DBU/h

At the bottom, there's a taskbar with various icons and system status indicators.

A online-retails-azure-databricks- x Compute - Databricks x +

adb-2702396966592863.3.azuredatabricks.net/compute?o=2702396966592863

Microsoft Azure | databricks

Search data, notebooks, recents, and more... CTRL + P

online-retails-azure-databricks-premium 2

+ New

Workspace

Recents

Catalog

Workflows

Compute

Marketplace

SQL

SQL Editor

Queries

Dashboards

Genie

Alerts

Query History

SQL Warehouses

Data Engineering

Job Runs

Data Ingestion

Pipelines

Machine Learning

Playground

Compute

All-purpose compute Job compute SQL warehouses Vector Search Pools Policies

Filter compute you have access to Created by Only pinned Create with Personal Compute

State	Name	Policy	Runtime	Active mem...	Active cores	Active DBU ...	Source	Creator	Notebooks
Running	Olist Data Processing Cluster	-	15.4	-	-	-	UI	20EM1A6117 su...	...

Previous Next 20 / page 18:51 05-04-2025 ENG IN

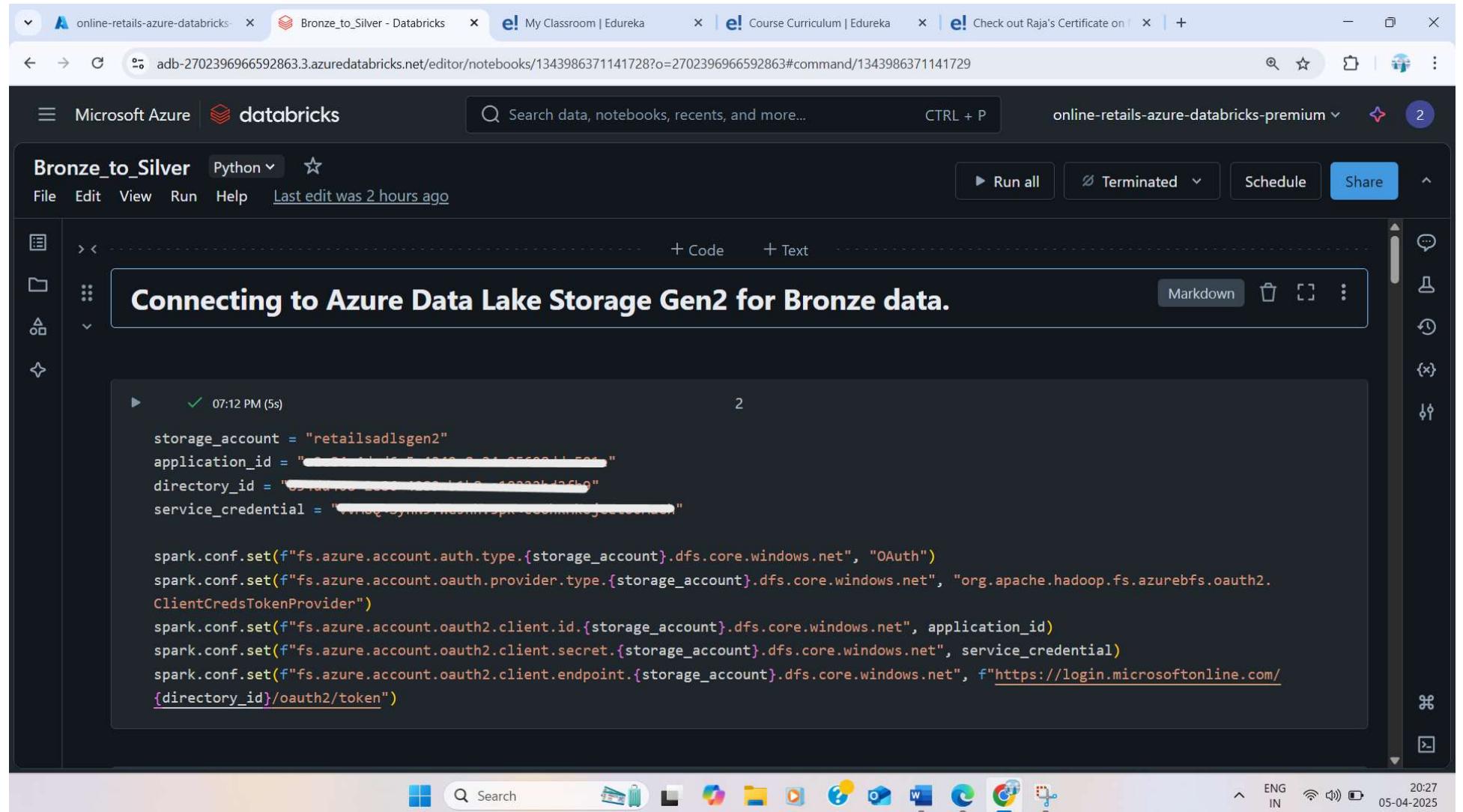
The screenshot shows the Microsoft Azure Databricks Compute page. On the left, there's a sidebar with various navigation options like Workspace, Recents, Catalog, Workflows, Compute (which is selected), Marketplace, SQL, SQL Editor, Queries, Dashboards, Genie, Alerts, Query History, SQL Warehouses, Data Engineering, Job Runs, Data Ingestion, Pipelines, Machine Learning, and Playground. The main area is titled 'Compute' and has tabs for All-purpose compute, Job compute, SQL warehouses, Vector Search, Pools, and Policies. It features a search bar and filters for 'Created by' and 'Only pinned'. A table lists the clusters, showing columns for State, Name, Policy, Runtime, Active memory, Active cores, Active DBUs, Source, Creator, and Notebooks. One cluster, 'Olist Data Processing Cluster', is listed with a runtime of 15.4. At the bottom, there are links for Previous, Next, and page 20, along with system status icons for battery, signal, and date/time.

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 explanation:-

Connecting to Azure Data Lake Storage Gen2 for Bronze data.



The screenshot shows a Databricks notebook interface. The title bar has tabs for 'online-retails-azure-databricks-' (active), 'Bronze_to_Silver - Databricks', 'My Classroom | Edureka', 'Course Curriculum | Edureka', 'Check out Raja's Certificate on', and a '+' button. The address bar shows the URL: <adb-2702396966592863.3.azuredatabricks.net/editor/notebooks/1343986371141728?o=2702396966592863#command/1343986371141729>. The notebook header includes 'Microsoft Azure' logo, 'databricks' logo, a search bar ('Search data, notebooks, recents, and more...'), 'CTRL + P', and a dropdown for 'online-retails-azure-databricks-premium'. Below the header are buttons for 'Run all', 'Terminated', 'Schedule', and 'Share'. The notebook itself has a title 'Bronze_to_Silver' with a Python icon and a star. The file menu includes 'File', 'Edit', 'View', 'Run', 'Help', and a note 'Last edit was 2 hours ago'. The main area contains a code cell with the following content:

```
storage_account = "retailsadlsgen2"
application_id = "REDACTED"
directory_id = "REDACTED"
service_credential = "REDACTED"

spark.conf.set(f"fs.azure.account.auth.type.{storage_account}.dfs.core.windows.net", "OAuth")
spark.conf.set(f"fs.azure.account.oauth.provider.type.{storage_account}.dfs.core.windows.net", "org.apache.hadoop.fs.azurebfs.oauth2.ClientCredsTokenProvider")
spark.conf.set(f"fs.azure.account.oauth2.client.id.{storage_account}.dfs.core.windows.net", application_id)
spark.conf.set(f"fs.azure.account.oauth2.client.secret.{storage_account}.dfs.core.windows.net", service_credential)
spark.conf.set(f"fs.azure.account.oauth2.client.endpoint.{storage_account}.dfs.core.windows.net", f"https://login.microsoftonline.com/{directory_id}/oauth2/token")
```

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.

```
storage_account = "retailsadlsgen2"
container_name = "olist"
folder_name = "Bronze"
base_path = f"abfss://{{container_name}}@{{storage_account}}.dfs.core.windows.net/{{folder_name}}/"
orders_path = base_path + "olist_orders_dataset.csv"
payments_path = base_path + "olist_order_payments_dataset.csv"
reviews_path = base_path + "olist_order_reviews_dataset.csv"
items_path = base_path + "olist_order_items_dataset.csv"
customers_path = base_path + "olist_customers_dataset.csv"
geolocation_path = base_path + "olist_geolocation_dataset.csv"
products_path = base_path + "olist_products_dataset.csv"
sellers_path = base_path + "olist_sellers_dataset.csv"

orders_df = spark.read.format("csv").option("header", "true").load(orders_path)
payments_df = spark.read.format("csv").option("header", "true").load(payments_path)
reviews_df = spark.read.format("csv").option("header", "true").load(reviews_path)
items_df = spark.read.format("csv").option("header", "true").load(items_path)
customers_df = spark.read.format("csv").option("header", "true").load(customers_path)
geolocation_df = spark.read.format("csv").option("header", "true").load(geolocation_path)
products_df = spark.read.format("csv").option("header", "true").load(products_path)
sellers_df = spark.read.format("csv").option("header", "true").load(sellers_path)

▶ (8) Spark Jobs
▶ [ ] customers_df: pyspark.sql.dataframe.DataFrame = [customer_id: string, customer_unique_id: string ... 3 more fields]
▶ [ ] geolocation_df: pyspark.sql.dataframe.DataFrame = [geolocation_zip_code_prefix: string, geolocation_lat: string ... 3 more fields]
▶ [ ] items_df: pyspark.sql.dataframe.DataFrame = [order_id: string, order_item_id: string ... 5 more fields]
▶ [ ] orders_df: pyspark.sql.dataframe.DataFrame = [order_id: string, customer_id: string ... 6 more fields]
▶ [ ] payments_df: pyspark.sql.dataframe.DataFrame = [order_id: string, payment_sequential: string ... 3 more fields]
▶ [ ] products_df: pyspark.sql.dataframe.DataFrame = [product_id: string, product_category_name: string ... 7 more fields]
▶ [ ] reviews_df: pyspark.sql.dataframe.DataFrame = [review_id: string, order_id: string ... 5 more fields]
▶ [ ] sellers_df: pyspark.sql.dataframe.DataFrame = [seller_id: string, seller_zip_code_prefix: string ... 2 more fields]
```

Remove duplicate and Null values :-

Cleaning the data

```
▶ ✓ 07:12 PM (<1s) 10
```

```
from pyspark.sql.functions import col, to_date, datediff, current_date, when
```

```
▶ ✓ 07:12 PM (<1s) 11
```

```
def clean_dataframe(df,name):  
    print("Cleaning "+ name)  
    return df.dropDuplicates().na.drop('all')  
orders_df = clean_dataframe(orders_df,"Orders")
```

```
▶ 📄 orders_df: pyspark.sql.dataframe.DataFrame = [order_id: string, customer_id: string ... 6 more fields]
```

```
Cleaning Orders
```

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

```
▶ ✓ 20 hours ago (<1s) 12
```

```
orders_df = orders_df.withColumn('order_purchase_timestamp',to_date(col('order_purchase_timestamp')))\n    .withColumn('order_delivered_customer_date',to_date(col('order_delivered_customer_date')))\n    .withColumn('order_estimated_delivery_date',to_date(col('order_estimated_delivery_date')))
```

```
▶ 📄 orders_df: pyspark.sql.dataframe.DataFrame = [order_id: string, customer_id: string ... 6 more fields]
```

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

▶ ✓ 20 hours ago (<1s)

13

```
orders_df = orders_df.withColumn('actual_delivery_time',datediff(col('order_delivered_customer_date'),col('order_purchase_timestamp')))  
orders_df = orders_df.withColumn('estimated_delivery_time',datediff(col('order_estimated_delivery_date'),col('order_purchase_timestamp')))  
orders_df = orders_df.withColumn('delay',when(col('actual_delivery_time') > col('estimated_delivery_time'),True).otherwise(False))  
orders_df = orders_df.withcolumn('delay_time',col('actual_delivery_time')-col('estimated_delivery_time'))  
  
▶ [orders_df: pyspark.sql.dataframe.DataFrame = [order_id: string, customer_id: string ... 10 more fields]]
```

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

▶ ✓ 20 hours ago (<1s)

16

```
orders_customers_df = orders_df.join(customers_df,['customer_id'],'left')  
orders_payments_df = orders_customers_df.join(payments_df,['order_id'],'left')  
orders_items_df = orders_payments_df.join(items_df, 'order_id', 'left')  
order_items_products_df = orders_items_df.join(products_df, ['product_id'] , 'left')  
final_df = order_items_products_df.join(sellers_df, ['seller_id'], 'left')  
  
▶ [final_df: pyspark.sql.dataframe.DataFrame = [seller_id: string, product_id: string ... 35 more fields]]  
▶ [order_items_products_df: pyspark.sql.dataframe.DataFrame = [product_id: string, order_id: string ... 32 more fields]]  
▶ [orders_customers_df: pyspark.sql.dataframe.DataFrame = [customer_id: string, order_id: string ... 14 more fields]]  
▶ [orders_items_df: pyspark.sql.dataframe.DataFrame = [order_id: string, customer_id: string ... 24 more fields]]  
▶ [orders_payments_df: pyspark.sql.dataframe.DataFrame = [order_id: string, customer_id: string ... 18 more fields]]
```

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 = "XXXXXXXXXX"
database = "XXXXXXXXXX"
port = "27018"
username = "olistnosqldb_mainlytoy"
password = "XXXXXXXXXXXXXX"

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.

▶ ✓ 20 hours ago (<1s)

17

```
final_df = final_df.join(mongo_data, "product_category_name", "left")
▶ final_df: pyspark.sql.dataframe.DataFrame = [product_category_name: string, seller_id: string ... 36 more fields]
```

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.

▶ ✓ 20 hours ago (22s)

18

```
final_df.write.mode('overwrite').parquet("abfss://olist@retailsadlsgen2.dfs.core.windows.net/silver")
▶ (8) Spark Jobs
```

This is My 'Silver' folder

The screenshot shows the Microsoft Azure Storage Container Overview page for the 'olist' container. The container name is 'olist / Silver'. The table lists 12 blobs, all of which are 'committed' type. The columns include Name, Modified, Access tier, Archive status, Blob type, Size, and Lease state. The blobs are timestamped between April 5, 2025, and April 23, 2025.

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
[..]						...
_committed_2518561937330943705	3/23/2025, 5:32:38 PM	Hot (Inferred)		Block blob	811 B	Available
_committed_3364807418193261340	4/5/2025, 7:19:59 PM	Hot (Inferred)		Block blob	815 B	Available
_committed_37182999355065541	3/23/2025, 4:30:19 PM	Hot (Inferred)		Block blob	1.6 KiB	Available
_committed_429288959890638407	3/23/2025, 5:29:25 PM	Hot (Inferred)		Block blob	1.19 KiB	Available
_committed_vacuum2974842127922467855	4/5/2025, 7:20:00 PM	Hot (Inferred)		Block blob	161 B	Available
_started_3364807418193261340	4/5/2025, 7:19:51 PM	Hot (Inferred)		Block blob	0 B	Available
part-00000-tid-3364807418193261340-9c91d991...	4/5/2025, 7:19:59 PM	Hot (Inferred)		Block blob	4.89 MiB	Available
part-00001-tid-3364807418193261340-9c91d991...	4/5/2025, 7:19:59 PM	Hot (Inferred)		Block blob	4.9 MiB	Available
part-00002-tid-3364807418193261340-9c91d991...	4/5/2025, 7:19:59 PM	Hot (Inferred)		Block blob	4.91 MiB	Available
part-00003-tid-3364807418193261340-9c91d991...	4/5/2025, 7:19:59 PM	Hot (Inferred)		Block blob	4.91 MiB	Available

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 .

A retail-synapse-workspace - Microsoft Azure

portal.azure.com/#@rajasuhashkesarigmail.onmicrosoft.com/resource/subscriptions/394a429d-397d-4207-af88-ca3b078d5e5a/resourceGroups/Data-Driven-Decision-Making-for-Enhanced-Online-Retail-Experience-and-Increased-Sales

Microsoft Azure Search resources, services, and docs (G+/-) Copilot Home Data-Driven-Decision-Making-for-Enhanced-Online-Retail-Experience-and-Increased-Sales retail-synapse-workspace Synapse workspace

Search New dedicated SQL pool New Apache Spark pool Refresh Reset SQL admin password Delete

Overview JSON View

Essentials

Resource group (move)	: Data-Driven-Decision-Making-for-Enhanced-Online-Retail-Experi...	Networking	: Show firewall settings
Status	: Succeeded	Primary ADLS Gen2 acco...	: https://retailsadlsgen2.dfs.core.windows.net
Location	: Central India	Primary ADLS Gen2 file s...	: retailsfilesystem
Subscription (move)	: Pay-As-You-Go	SQL admin username	: sqladminuser
Subscription ID	: 394a429d-397d-4207-af88-ca3b078d5e5a	SQL Microsoft Entra admin	: live.com#rajasuhashkesari@gmail.com
Managed virtual network	: No	Dedicated SQL endpoint	: retails-synapse-workspace.sql.azuresynapse.net
Managed Identity object ...	: 7b06390f-c952-45a2-af29-8b653538f3b7	Serverless SQL endpoint	: retails-synapse-workspace-on-demand.sql.azuresynapse.net
Workspace web URL	: https://web.azuresynapse.net?workspace=%2bsubscriptions%2f39...	Development endpoint	: https://retails-synapse-workspace.dev.azuresynapse.net
Tags (edit)	: Add tags		

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

A retail-synapse-workspace - Mi 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 > retail-synapse-workspace Microsoft Azure | Synapse Analytics > retail-synapse-workspace

retails-synapse-workspace

Synapse Analytics workspace

New ▾

Ingest

Explore and analyze

Visualize

Discover more

Knowledge center

Browse partners

Recent resources

Name

Last opened by you

Search

Windows Start button

Icons for various applications: File Explorer, Task View, Control Panel, File History, OneDrive, Google Chrome, Microsoft Edge, and others.

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

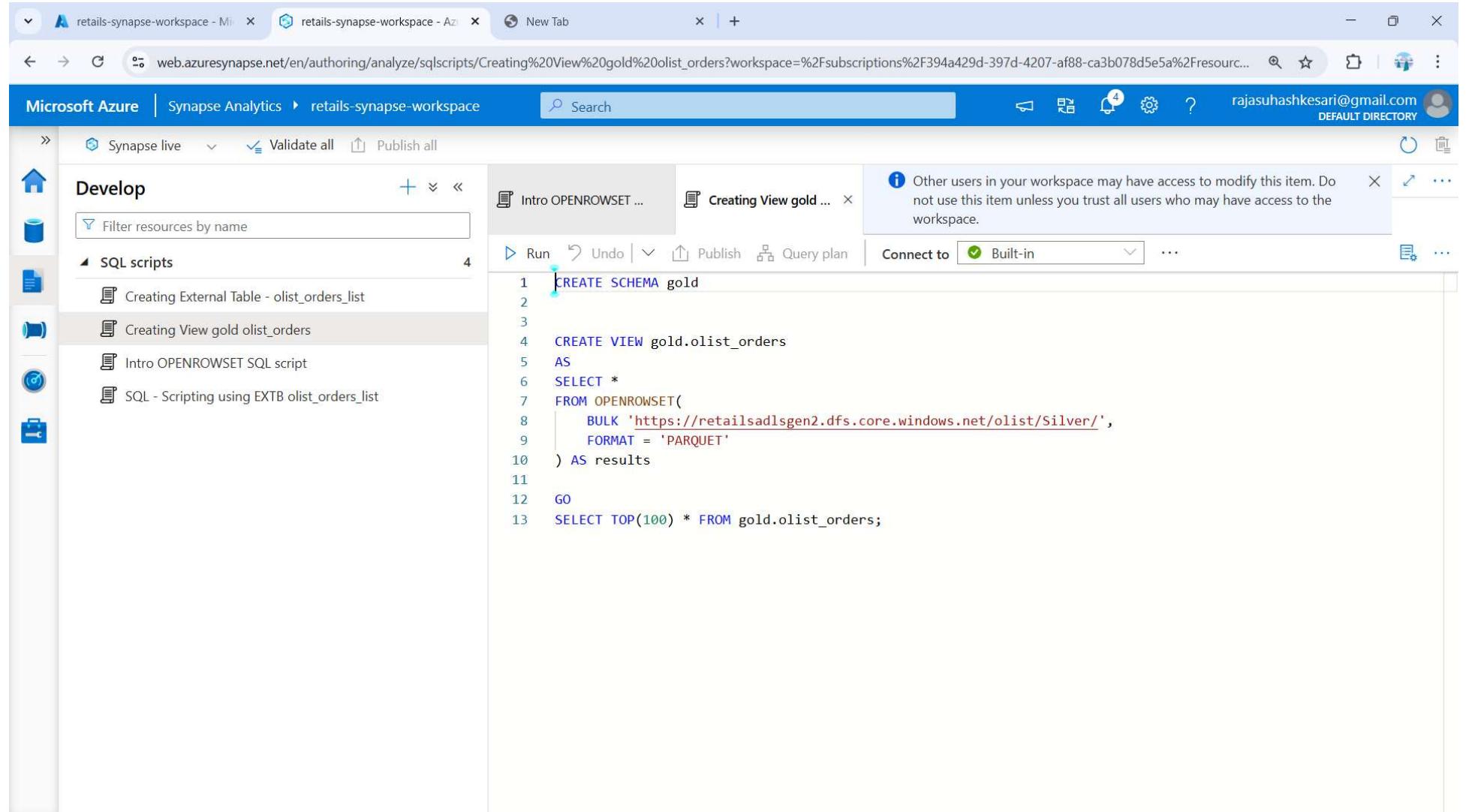
Example:-

```
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'

The screenshot shows the Microsoft Azure Synapse Analytics Data workspace interface. The top navigation bar includes 'Microsoft Azure' and the workspace name 'retails-synapse-workspace'. Below the navigation bar, there are buttons for 'Synapse live', 'Validate all', and 'Publish all'. The left sidebar features icons for Home, Database, Table, View, and Security. The main area is titled 'Data' and has tabs for 'Workspace' (selected) and 'Linked'. A search bar at the top of the main area contains the placeholder 'Filter resources by name'. Under the 'Workspace' tab, a section for 'SQL database' is expanded, showing one item named 'olist (SQL)' which is further expanded to show 'External tables', 'External resources', 'Views', 'Schemas', and 'Security'.

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



The screenshot shows the Microsoft Azure Synapse Analytics workspace interface. The left sidebar is titled "Develop" and lists several SQL scripts under the "SQL scripts" section, including "Creating External Table - olist_orders_list", "Creating View gold olist_orders", "Intro OPENROWSET SQL script", and "SQL - Scripting using EXTB olist_orders_list". The main area displays a SQL script titled "Creating View gold ...". The script creates a schema "gold" and a view "gold.olist_orders" that reads data from a BULK-IMPORTed table in a Silver layer storage account. A tooltip message at the top right of the script editor states: "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." The script itself is as follows:

```
1 CREATE SCHEMA gold
2
3
4 CREATE VIEW gold.olist_orders
5 AS
6 SELECT *
7 FROM OPENROWSET(
8     BULK 'https://retailsadlsgen2.dfs.core.windows.net/olist/Silver/',
9     FORMAT = 'PARQUET'
10 ) AS results
11
12 GO
13 SELECT TOP(100) * FROM gold.olist_orders;
```

A retails-synapse-workspace - Mi x A retails-synapse-workspace - Az x New Tab x | +

← → ⌂ web.azuresynapse.net/en/authoring/explore/workspace?workspace=%2Fsubscriptions%2F394a429d-397d-4207-af88-ca3b078d5e5a%2FresourceGroups%2FData-Driven-Decision-Making... Search 4 notifications ? rajasuhashkesari@gmail.com DEFAULT DIRECTORY

Microsoft Azure | Synapse Analytics > retails-synapse-workspace

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

+

Search

Cloud

Help

?

rajasuhashkesari@gmail.com

DEFAULT DIRECTORY

↻

⟳

SQL database 1

olist (SQL)

External tables

External resources

Views

gold.olist_orders

System views

Schemas

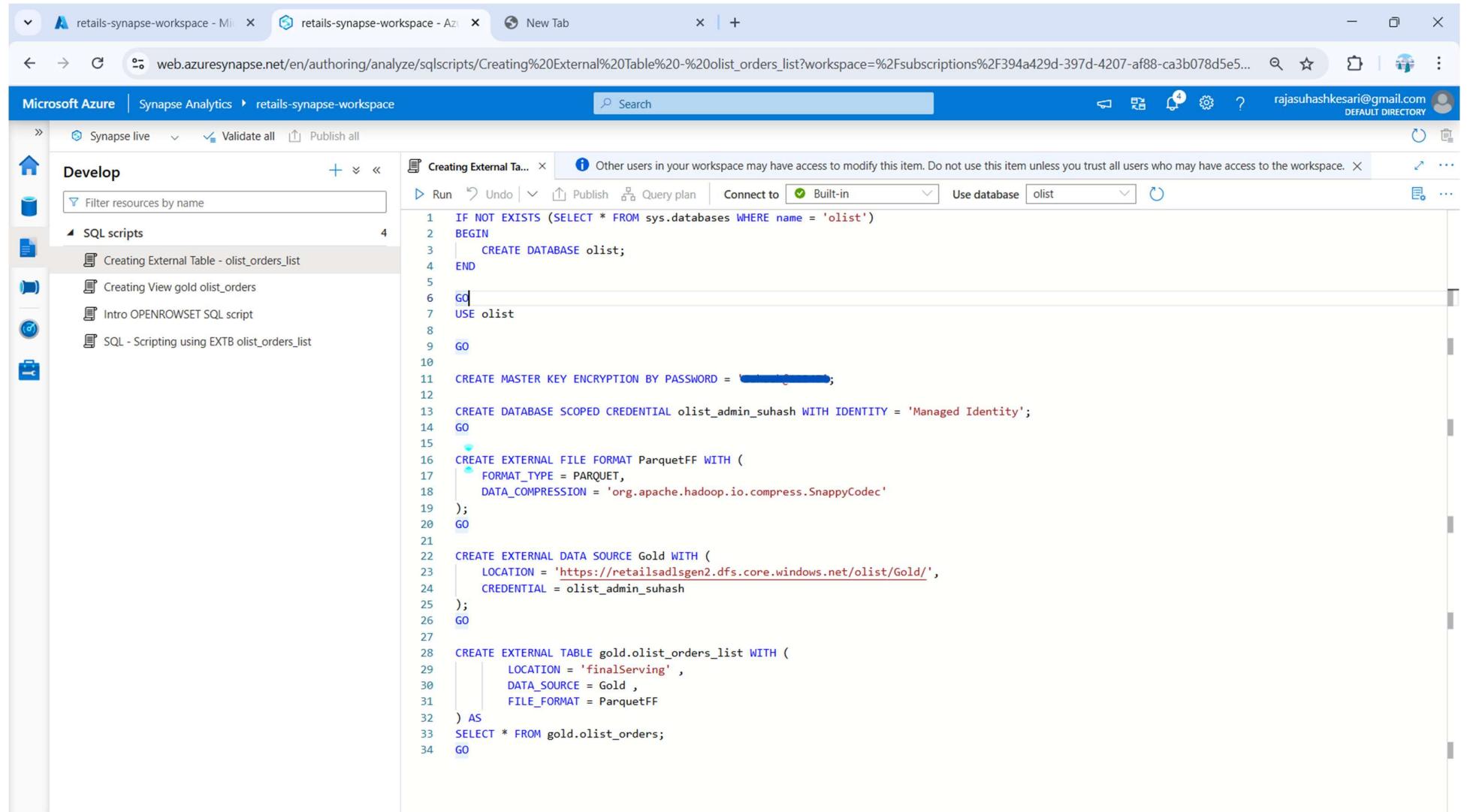
Security

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 .



The screenshot shows the Microsoft Azure Synapse Analytics workspace interface. The left sidebar is titled 'Develop' and shows a list of SQL scripts: 'Creating External Table - olist_orders_list', 'Creating View gold olist_orders', 'Intro OPENROWSET SQL script', and 'SQL - Scripting using EXTB olist_orders_list'. The main area is titled 'Creating External Ta...' and contains the following SQL script:

```
1 IF NOT EXISTS (SELECT * FROM sys.databases WHERE name = 'olist')
2 BEGIN
3     CREATE DATABASE olist;
4 END
5
6 GO
7 USE olist
8
9 GO
10
11 CREATE MASTER KEY ENCRYPTION BY PASSWORD = 'XXXXXXXXXX';
12
13 CREATE DATABASE SCOPED CREDENTIAL olist_admin_suhash WITH IDENTITY = 'Managed Identity';
14 GO
15
16 CREATE EXTERNAL FILE FORMAT ParquetFF WITH (
17     FORMAT_TYPE = PARQUET,
18     DATA_COMPRESSION = 'org.apache.hadoop.io.compress.SnappyCodec'
19 );
20 GO
21
22 CREATE EXTERNAL DATA SOURCE Gold WITH (
23     LOCATION = 'https://retailsadlsgen2.dfs.core.windows.net/olist/Gold/',
24     CREDENTIAL = olist_admin_suhash
25 );
26 GO
27
28 CREATE EXTERNAL TABLE gold.olist_orders_list WITH (
29     LOCATION = 'finalServing' ,
30     DATA_SOURCE = Gold ,
31     FILE_FORMAT = ParquetFF
32 ) AS
33 SELECT * FROM gold.olist_orders;
34 GO
```

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

The screenshot shows the Microsoft Azure Synapse Analytics Data workspace interface. The top navigation bar includes 'Microsoft Azure' and 'Synapse Analytics' with a dropdown for 'retails-sy'. Below the navigation is a toolbar with icons for 'Synapse live', 'Validate all', and a refresh button. The main area is titled 'Data' with tabs for 'Workspace' (selected) and 'Linked'. A search bar at the top says 'Filter resources by name'. Under 'Workspace', there is a tree view of resources:

- SQL database** (1 item):
 - olist (SQL)**:
 - External tables**:
 - gold.olist_orders_list**:
 - Columns**
 - External resources**:
 - External data sources**
 - External file formats**
 - Views**
 - Schemas**
 - Security**

After creating external table I got data from silver to gold

The screenshot shows the Microsoft Azure Storage Container Overview page for the 'olist' container. The container name is 'olist'. The 'Overview' tab is selected. The page displays blob data with the following columns: Name, Modified, Access tier, Archive status, Blob type, Size, and Lease state. There are six blobs listed:

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
[...]	3/24/2025, 6:55:53 PM	Hot (Inferred)		Block blob	0 B	Available
_	3/24/2025, 6:55:59 PM	Hot (Inferred)		Block blob	4.91 MiB	Available
28986283-5966-4B3E-A757-E41B31F9927F_8_0-1....	3/24/2025, 6:56:00 PM	Hot (Inferred)		Block blob	4.9 MiB	Available
28986283-5966-4B3E-A757-E41B31F9927F_8_0-16...	3/24/2025, 6:56:00 PM	Hot (Inferred)		Block blob	4.92 MiB	Available
28986283-5966-4B3E-A757-E41B31F9927F_8_0-4....	3/24/2025, 6:56:00 PM	Hot (Inferred)		Block blob	4.92 MiB	Available
28986283-5966-4B3E-A757-E41B31F9927F_8_0-7....	3/24/2025, 6:56:00 PM	Hot (Inferred)		Block blob	4.92 MiB	Available

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

retails-synapse-workspace - Microsoft Edge retails-synapse-workspace - Azure 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 > retail-synapse-workspace

Search

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 delivered.
11 SELECT COUNT((order_item_id)) AS "Number of items Delivered" 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...	e6c386cf321a...	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 - Microsoft Edge retails-synapse-workspace - Azure 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 > retailssynapse-workspace

Search

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 delivered.
11 SELECT COUNT((order_item_id)) AS "Number of items Delivered" 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.

retails-synapse-workspace - Microsoft Edge retails-synapse-workspace - Azure 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 > retail-synapse-workspace

Search

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 delivered.
11 SELECT COUNT((order_item_id)) AS "Number of items Delivered" FROM gold.olist_orders_list WHERE order_status = 'Delivered';
12 GO
```

Results Messages

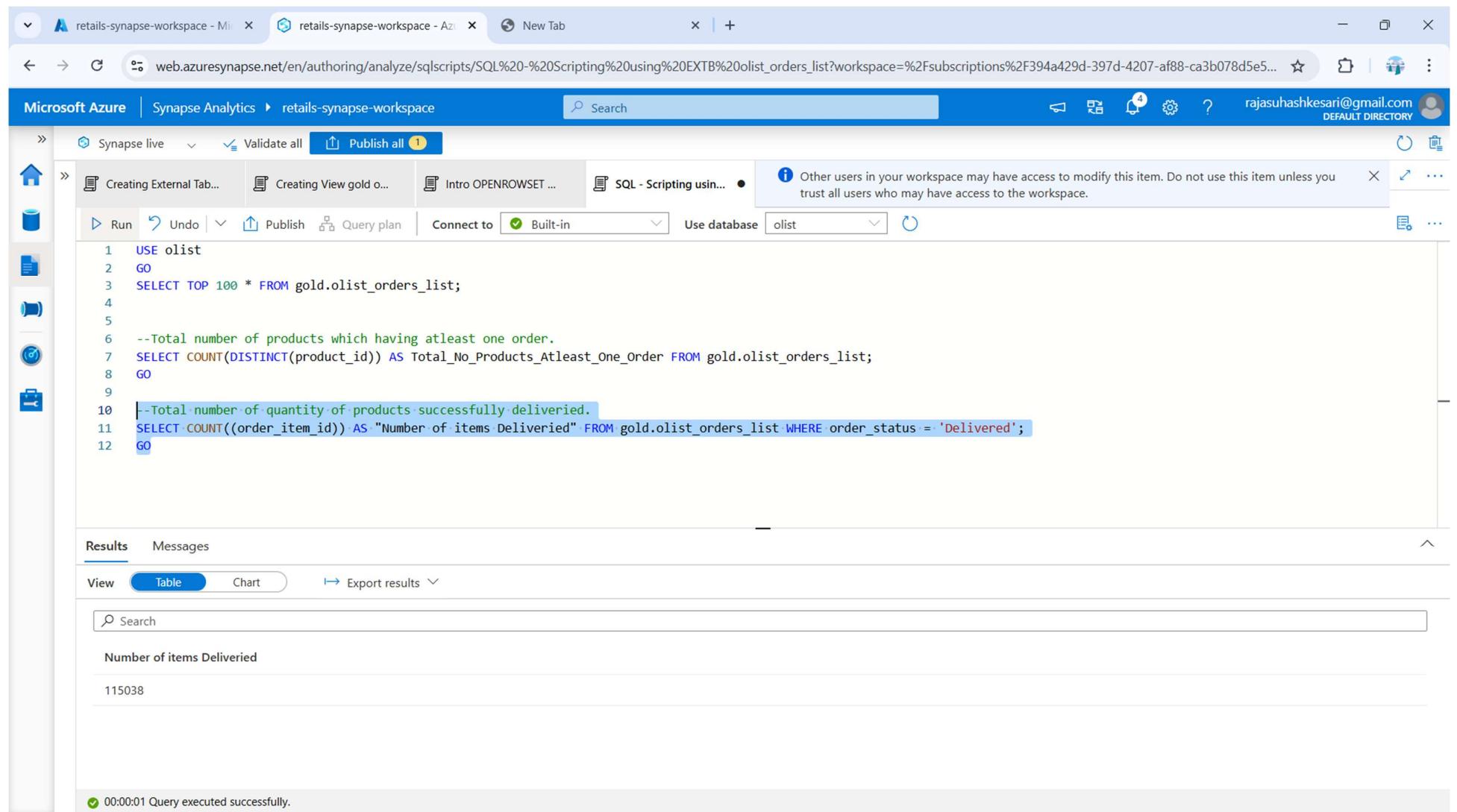
View Table Chart Export results

Search

Number of items Delivered

115038

00:00:01 Query executed successfully.



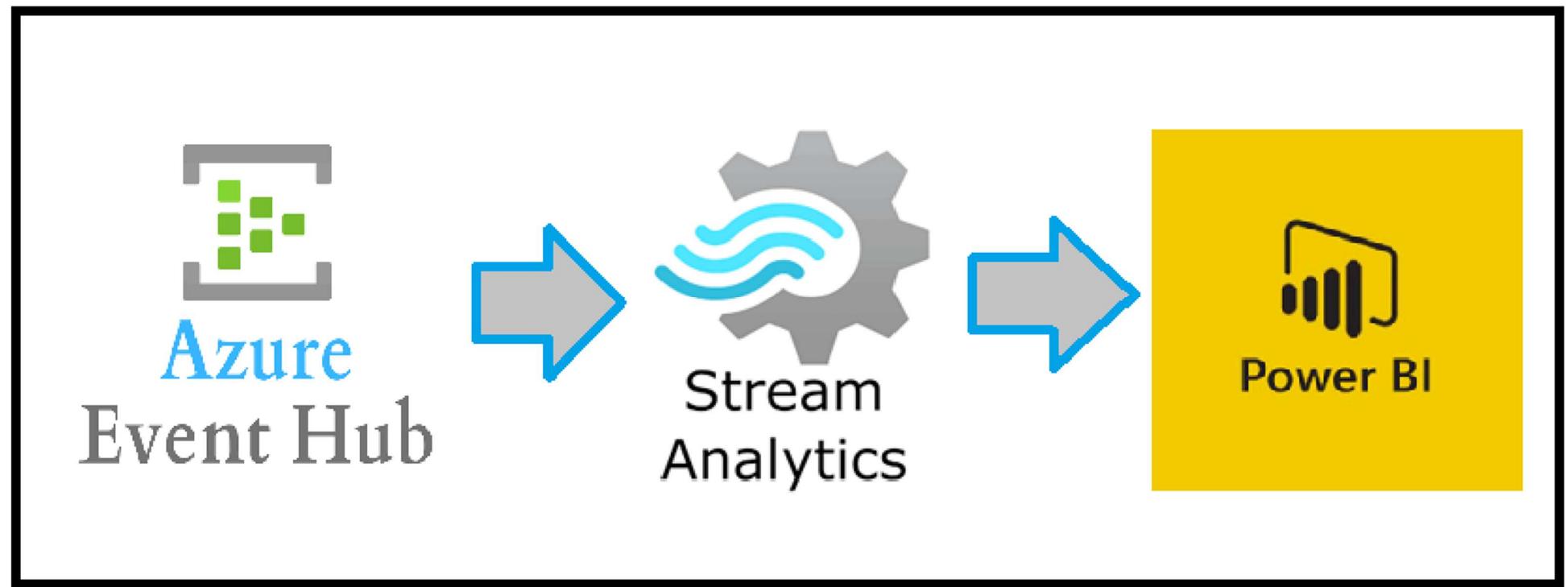
You see the outputs in the below of the picture.

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**

The screenshot shows the Microsoft Azure portal interface for the 'Olist-Retails-Real-Time-Data-Processing' resource group. The top navigation bar includes the Microsoft Azure logo, a search bar, and user information (rajasuhashkesarigmail.onmicrosoft.com). Below the header, the URL is portal.azure.com/#@rajasuhashkesarigmail.onmicrosoft.com/resource/subscriptions/394a429d-397d-4207-af88-ca3b078d5e5a/resourceGroups/Olist-Retails-Real-Time-Data-Processing/o... . The main content area displays the 'Overview' tab for the resource group. Key details shown include:

- Subscription (move) : Pay-As-You-Go
- Subscription ID : 394a429d-397d-4207-af88-ca3b078d5e5a
- Tags (edit) : Add tags
- Deployments : 1 Deploying, 2 Failed, 3 Succeeded
- Location : South India

The left sidebar lists navigation options: Overview, Activity log, Access control (IAM), Tags, Resource visualizer, Events, Settings, Cost Management, Monitoring, Automation, Help, Resources, and Recommendations. The 'Resources' tab is selected. A table lists resources:

Name ↑	Type ↑↓	Location ↑↓
Olist-Azure-Event-Hub-NameSpace	Event Hubs Namespace	South India
Olist-Retails-Stream-Analytics-Job	Stream Analytics job	South India

At the bottom, there are navigation buttons for < Previous, Page 1 of 1, Next >, and a Give feedback link.

Step - 1:- I have created 'Azure Event Hub' with name 'Olist-Azure-Event-Hub-NameSpace'.

Screenshot of the Microsoft Azure portal showing the 'Olist-Azure-Event-Hub-NameSpace' overview page.

The URL in the browser is: <https://portal.azure.com/#@rajasuhashkesarigmail.onmicrosoft.com/resource/subscriptions/394a429d-397d-4207-af88-ca3b078d5e5a/resourceGroups/Olist-Retails-Real-Time-Data-Processing/p...>

The page title is: Olist-Azure-Event-Hub-NameSpace

The left sidebar shows the following navigation items:

- Activity log
- Access control (IAM)
- Tags
- Diagnose and solve problems
- Data Explorer
- Resource visualizer
- Events
- Settings
- Entities
- Monitoring
- Automation
- Help

The main content area displays the 'Essentials' section with the following details:

Resource group (move) : Olist-Retails-Real-Time-Data-Processing		Created : Saturday, March 29, 2025 at 12:21:25 GMT+5:30
Status	: Active	Updated : Sunday, April 6, 2025 at 20:44:33 GMT+5:30
Location	: South India	Zone Redundancy : Not Enabled
Subscription (move)	: Pay-As-You-Go	Pricing tier : Basic
Subscription ID	: 394a429d-397d-4207-af88-ca3b078d5e5a	Throughput Units : 1 unit
Host name	: Olist-Azure-Event-Hub-NameSpace.servicebus.windows.net	Auto-inflate throughput ... : Not Supported
		Local Authentication : Enabled

Below the essentials section, there is a 'NAMESPACE CONTENTS' section showing '1 EVENT HUB' and a 'KAFKA SURFACE NOT SUPPORTED' message.

The time range for data visualization is set to '1 hour'.

The 'Requests' chart shows a single sharp peak reaching a value of 5.

The 'Messages' chart shows values from 10 to 100.

The 'Throughput' chart shows values from 10B to 100B.

Olist-Azure-Event-Hub-NameSpace

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 Event Hub Refresh Give feedback

Overview

Activity log

Access control (IAM)

Tags

Diagnose and solve problems

Data Explorer

Resource visualizer

Events

Settings

Entities

Event Hubs

olist-azure-event-hub

Status: Disabled

Message retention: 1 hour

Partition count: 1

Search to filter items by name...

Event Hubs 1

olist-azure-event-hub (Olist-Az...)

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 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!

Overview Access control (IAM) Diagnose and solve problems Data Explorer Resource visualizer Settings Entities Features Automation Help

Resource group ([move](#)) : [Olist-Retails-Real-Time-Data-Processing](#) Status : [Disabled](#)
Location : South India Namespace : [Olist-Azure-Event-Hub-NameSpace](#)
Subscription ([move](#)) : [Pay-As-You-Go](#) Created : Sunday, April 6, 2025 at 20:44:42 GMT+5:30
Subscription ID : 394a429d-397d-4207-af88-ca3b078d5e5a Updated : Sunday, April 6, 2025 at 20:48:01 GMT+5:30
Partition count : 1 Cleanup policy : [Delete](#)

JSON View

Essentials

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;SharedAccessKey=XXXXXXXXXX;EntityPath=olist-azure-event-hub"
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].items()}
        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': '7fbf3cba00dcf6da3f7c0fe6a162b4c1', 'order_id': '908e767e28908dc5ec61fe0ef8ef03f9', 'customer_id': '64d181ef52b1829800c598942459061b', 'customer_unique_id': 'd903ch2df3fc0df3e432hc359d8a6hf2', 'customer_zip_code_prefix': 37903, 'customer_city': 'passos', 'customer_state': 'MG', 'customer_full_name': 'maria', 'customer_email': 'maria@email.com', 'customer_phone': '(11) 98765-4321', 'customer_address': 'rua das rosas, 123', 'customer_neighborhood': 'centro', 'customer_longitude': -46.6333, 'customer_latitude': -23.5505, 'customer_is_frequent_buyer': false, 'customer_is_promotional_email_subscribed': true, 'customer_is_newsletter_subscribed': true, 'customer_is_reward_member': true, 'customer_mileage': 1200, 'customer_segment': 'silver', 'customer_is_delivery_available': true, 'customer_is_returning_customer': true, 'customer_is_verified': true, 'customer_is_promotional_email_subscribed': true, 'customer_is_newsletter_subscribed': true, 'customer_is_reward_member': true, 'customer_mileage': 1200, 'customer_segment': 'silver', 'customer_is_delivery_available': true, 'customer_is_returning_customer': true, 'customer_is_verified': true}
```

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

At azure event hub side we are receiving events below are the events received .

Screenshot of the Azure Event Hub Pipeline Data Explorer page for the 'olist-azure-event-hub' instance.

The page shows the following details:

- Event Hub:** olist-azure-event-hub
- Total received events:** 6
- Sequence Number:** 0 to 5
- Offset:** 0 to 3736
- Partition ID:** 0
- Enqueued Time:** Sun, Apr 06, 25, 10:25:23 PM GM...
- Content Type:** JSON objects representing product category data.
- Message ID:** Unique identifiers for each message.
- Event Body:** JSON objects containing product category names and seller IDs.

On the left sidebar, there are sections for sending events, inspecting properties (Partition ID set to 'All partition IDs', Consumer group set to '\$Default'), and setting the event position (Oldest position selected). There is also an 'Advanced properties' section and a 'View events' button.

At the bottom center, there is a magnifying glass icon over a document icon, likely indicating search or filtering functionality.

Azure Stream Analytics:-

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

The screenshot shows the Microsoft Azure portal interface for managing a Stream Analytics job named "Olist-Retails-Stream-Analytics-Job".

Job Overview:

- Status:** Starting
- Resource group:** Olist-Retails-Real-Time-Data-Processing
- Location:** South India
- Subscription:** Pay-As-You-Go
- Subscription ID:** 394a429d-397d-4207-af88-ca3b078d5e5a
- Pricing plan:** StandardV2 (manage)
- Tags:** Add tags

Monitoring: Monitoring tab is selected.

Troubleshooting:

- Errors and warnings:** The streaming job failed: Stream Analytics job has validation errors: The token has an invalid signature for the following EventHub... (Start job 'Olist-Retails-Stream...' 2 minutes ago)

Actions:

- Job diagram (preview)
- View all activity logs
- Enable diagnostics

`1 SELECT * INTO "streamto-powerbi" FROM "olist-azure-event-hub";`

product_category_name	seller_id	product_id	order_id	customer_id	customer_unique_id	customer_zip_code_prefix	customer_name
"papelaria"	"323ce52b5b81df2cd8..."	"8c1e9287e6b9508f46..."	"d82c5f84358d79a603..."	"dfc05ec4cd332bc05c..."	"c5f9b6b7f240e750a9..."	39980	"cachoeira"
"esporte_lazer"	"218d46b86c1881d02..."	"e44f675b60b3a3a245..."	"03fefb6f166efd081e0..."	"68c8855d1c948dff63..."	"26d571da801f51da17..."	5885	"sao paulo"
"pcs"	"a00824eb9093d40e5..."	"34f99d82fcf355d08d..."	"3a4b013e014723cc38..."	"e7c905bf4bb13543e8..."	"2c3b08cf3584d8c0a8..."	19023	"presidente"
"beleza_saude"	"db2956745b3a8e9f37..."	"9ab817d90bc61ed77..."	"49ac83265a3a9d7d49..."	"c47a4cc40c4b8ca461..."	"0a1e3cbe04dabc1a55..."	79570	"aparecid...
"perfumaria"	"7a425d299613df3e61..."	"e1f24e88490db8f0fd..."	"c0257fe04f54df9e09b..."	"bbe5971faf8ed849f7..."	"dd125e1d2a8cf2f724..."	36037	"juiz de fc"

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 :-

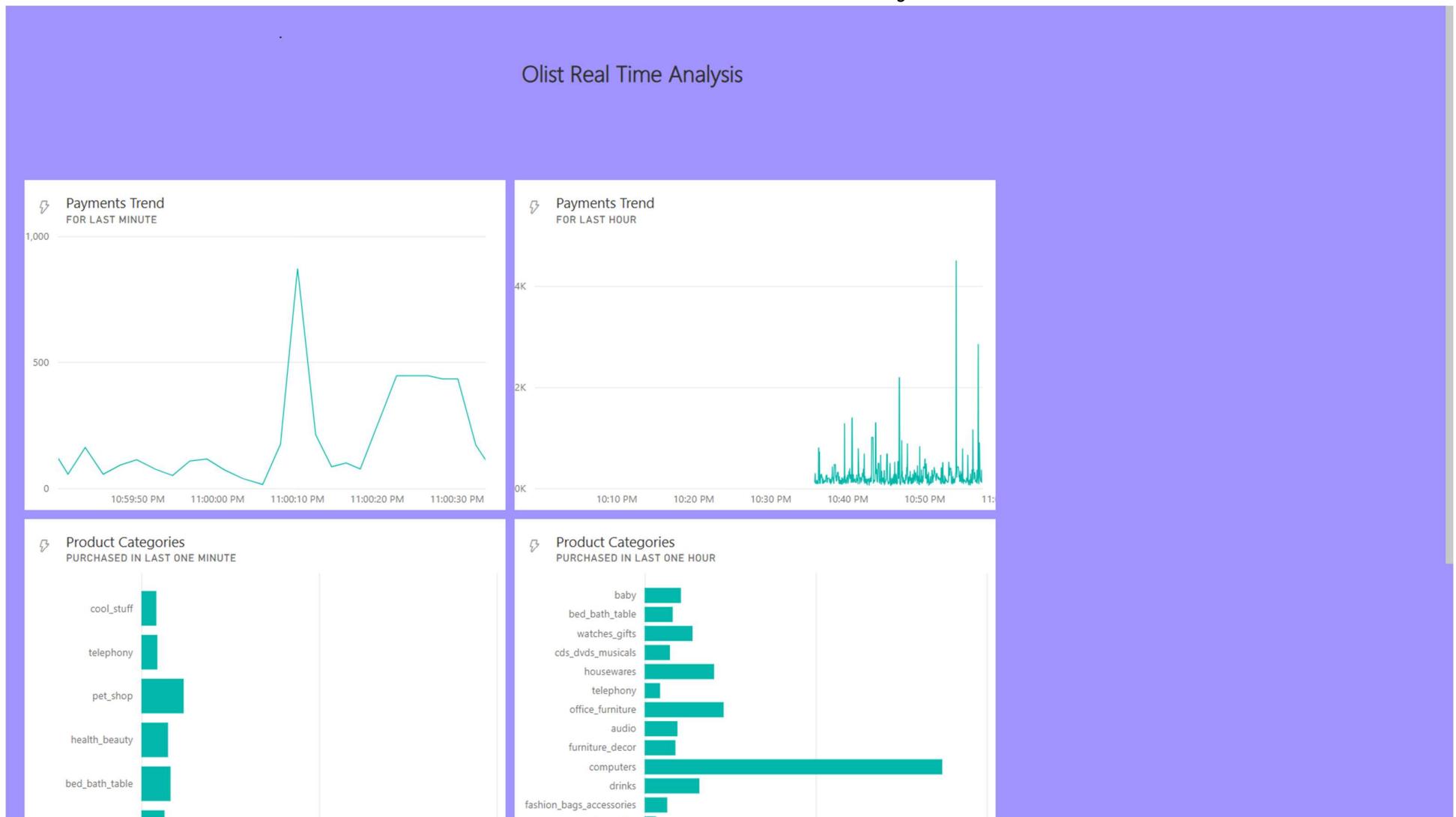
The screenshot shows the Microsoft Fabric workspace interface. The left sidebar lists various workspaces: Home, OneLake, Monitor, Real-Time, Workloads, and Olist-Real-Time-Data... (which is currently selected). The main area is titled "Olist-Real-Time-Data-Processing". It features a search bar, a trial status (45 days left), and links for Create deployment pipeline, Create app, Manage access, and Workspace settings. Below these are buttons for New item, New folder, and Import. A central section displays a table of task flows:

Name	Type	Task	Owner	Refreshed	Next refresh	Endorsement	Sensitivity	Included in app
Olist-Real-Time-Analysis	Dashboard	—	Olist-Real-Ti...	—	—	—	—	<input type="checkbox"/> No
Olist-streaming-Dataset	Semantic mo...	—	Olist-Real-Ti...	4/4/2025, 2:35...	N/A	—	—	

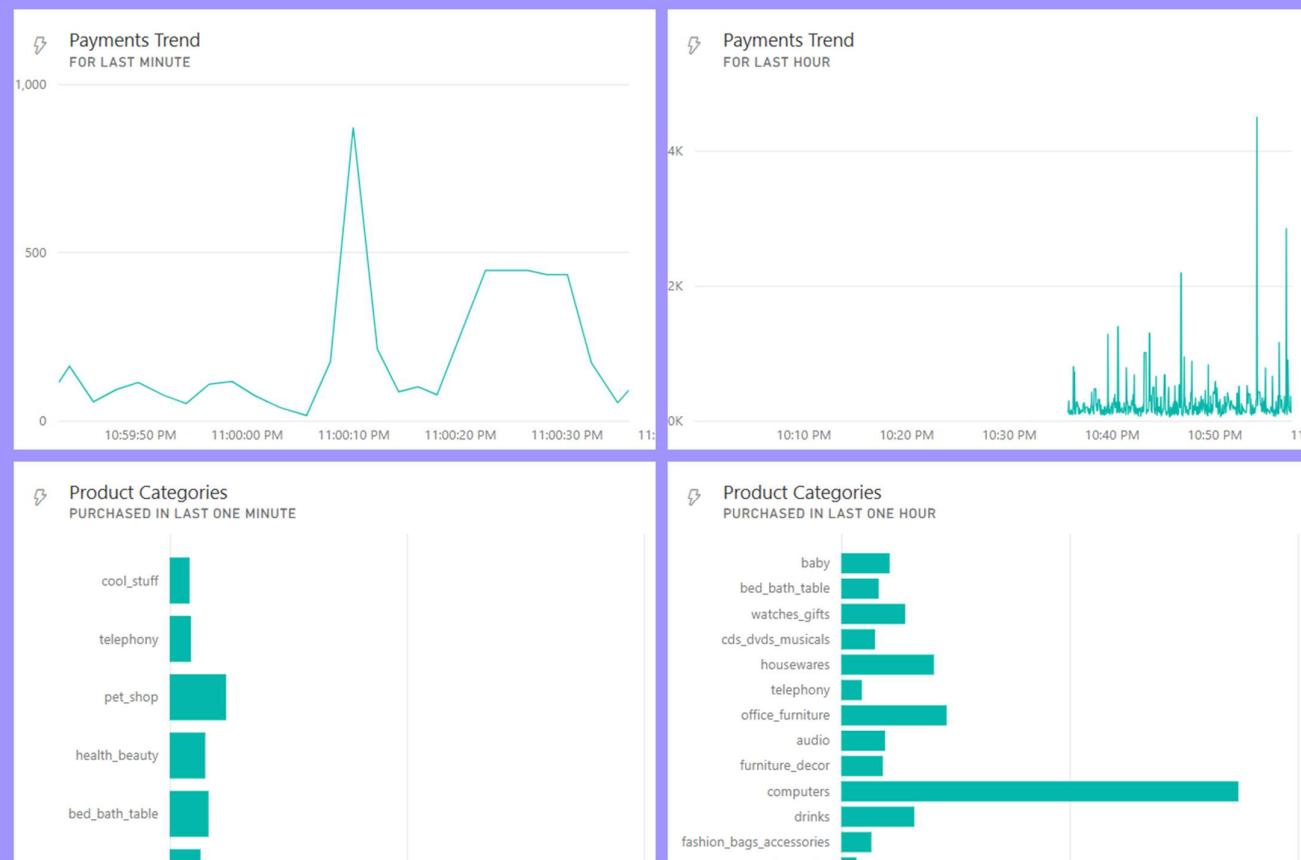
In the above Olist-streaming-Dataset is the semantic model which is receiving the Stream data.

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

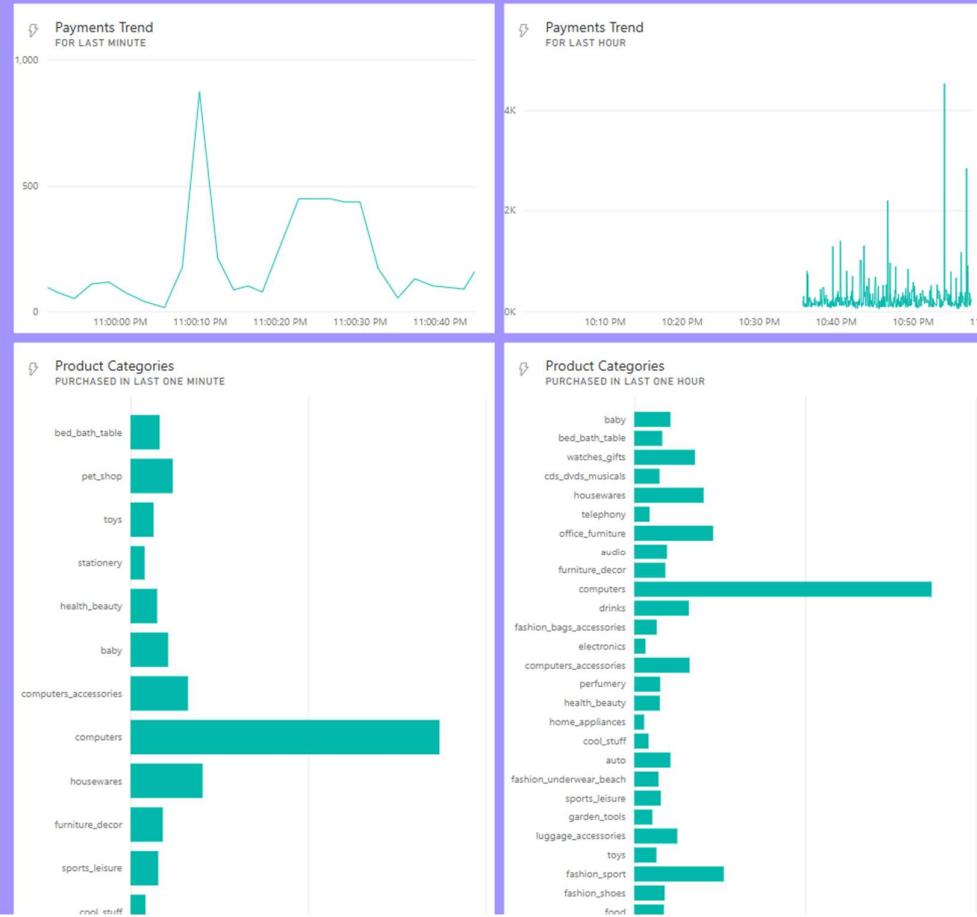
This is the live dashboard in which some tile will show real time analysis for last 1 minute and last 1 hour.



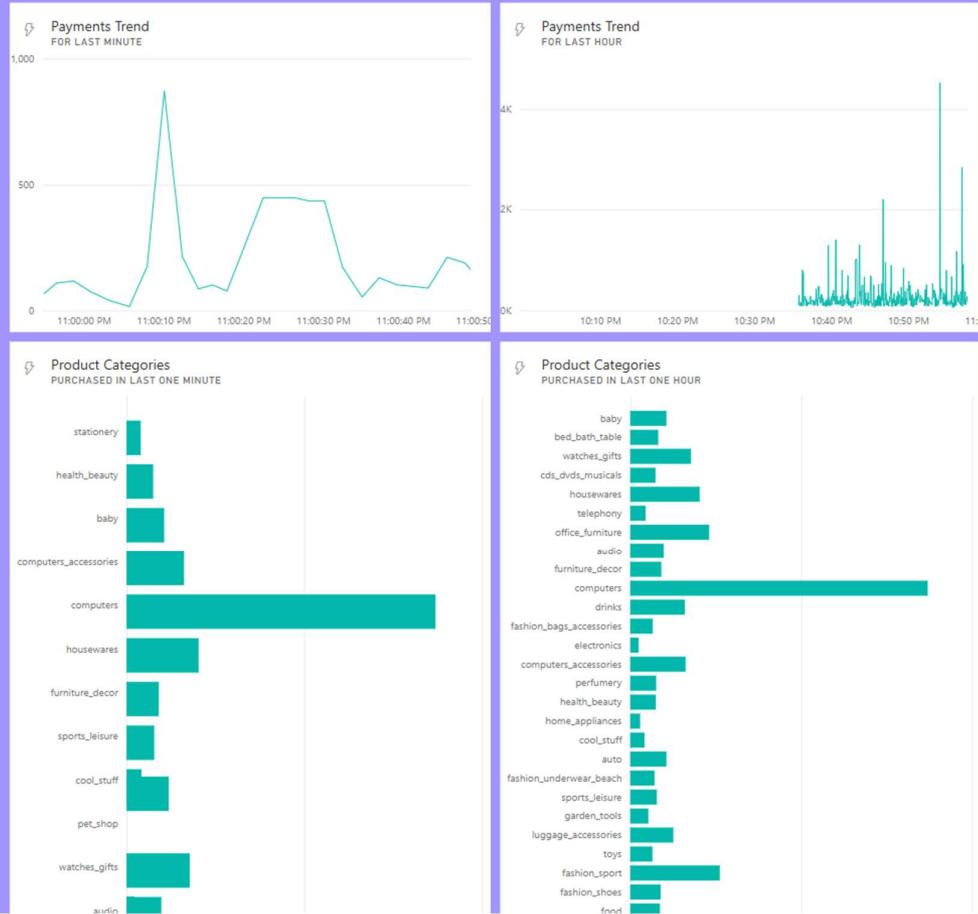
Olist Real Time Analysis



Olist Real Time Analysis



Olist Real Time Analysis



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

The End

Thank You