A complete End-to-End Azure Data Engineering project

Overview:

This project demonstrates a complete end-to-end data engineering pipeline using Azure cloud services with e-commerce data. The solution focuses on data ingestion, transformation, storage, and analytics using modern cloud-based tools.

Data used

E-Commerce Data Engineering Pipeline using Azure

Azure services used

1. Azure Data Lake Gen2 - For storing raw and processed data
2. Azure Databricks (Spark) - For data transformation and processing using PySpark
3. Azure Synapse Analytics - For querying and analysing the processed data
4. Azure Data factory – For ETL process and pipelines
5. Azure key vault - to securely store and manage secrets and certificates
6. Azure Entra ID - for securing and managing user identities and access to resources
7. MySQL and MongoDB – Databases

Project Objectives:

- Ingest raw e-commerce data into a data lake.
- Clean and transform data using PySpark in Azure Databricks.
- Store the cleaned data in structured format (Parquet/Delta) in ADLS.
- Analyze the processed data using Synapse SQL.

Pre-requisites:

Azure free-trail/paid subscription

https://azure.microsoft.com/en-us/pricing/purchase-options/azure-account

Service deployment documents:

https://learn.microsoft.com/en-us/azure/data-factory/quickstart-create-data-factory

https://learn.microsoft.com/en-us/azure/databricks/getting-started/

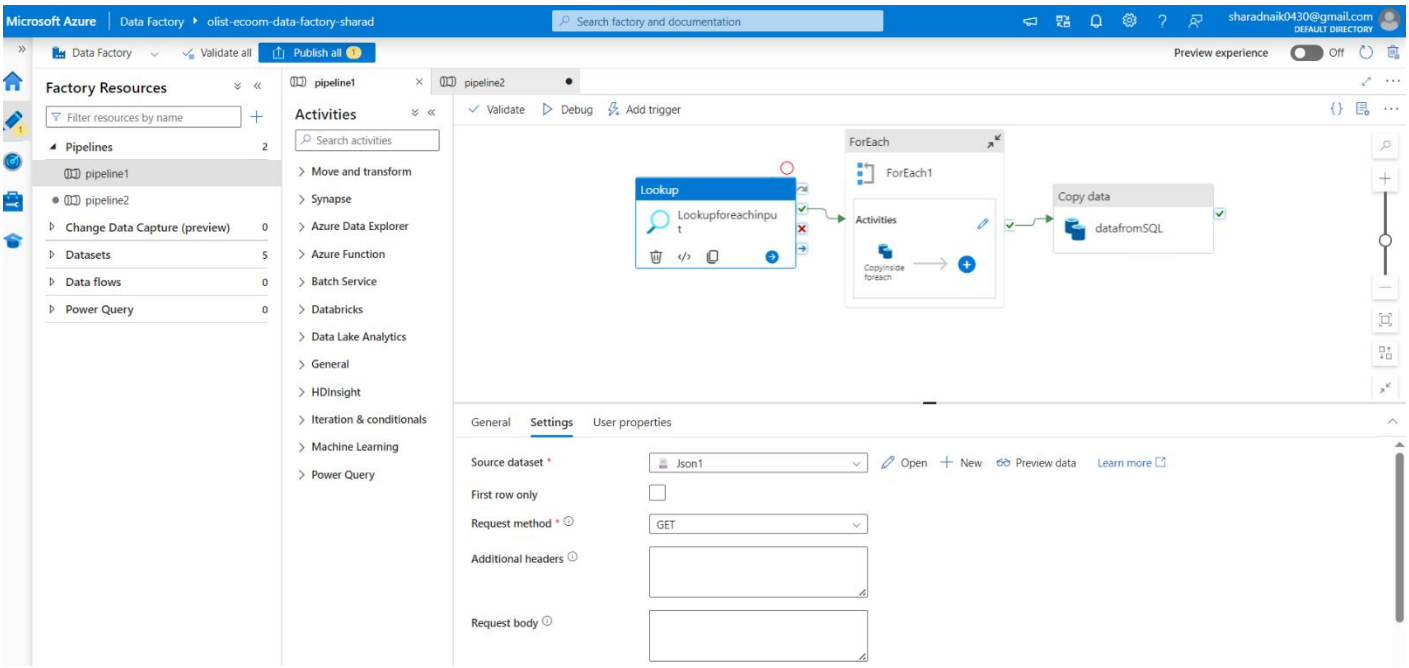
https://learn.microsoft.com/en-us/azure/storage/blobs/create-data-lake-storage-account

https://learn.microsoft.com/en-us/azure/synapse-analytics/get-started-create-workspace

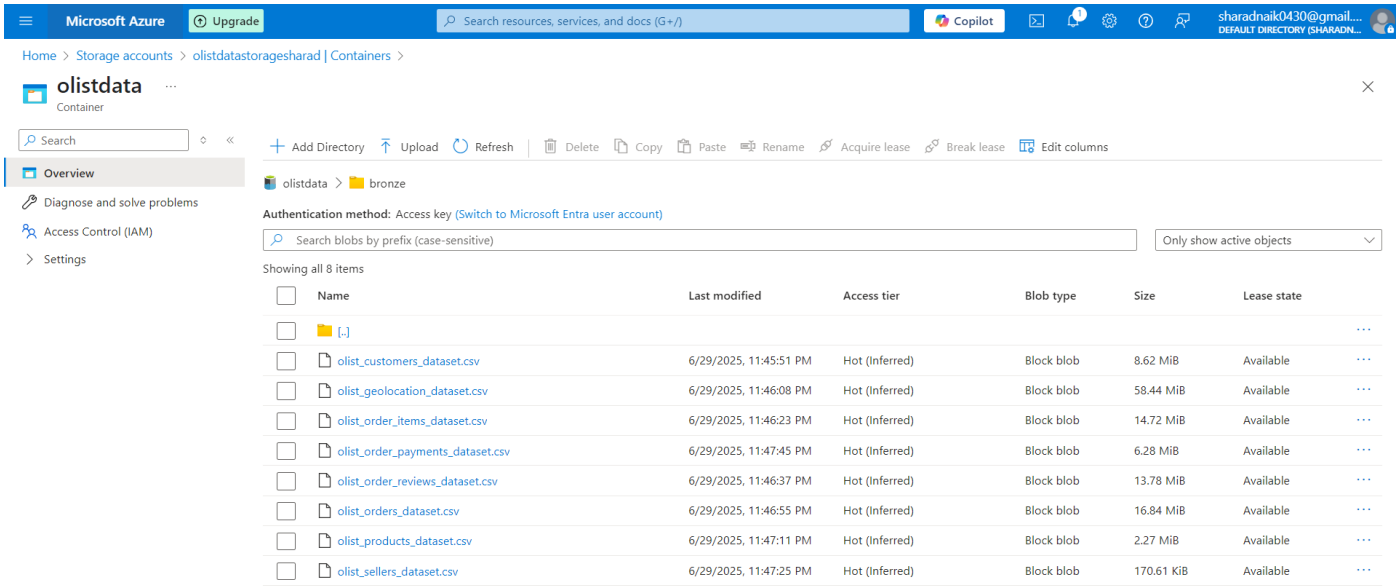
Workflow Overview:

Step 1: Data Extraction (Bronze Layer)

- Extracted raw data from a MySQL database using Azure Data Factory.
- Implemented parameterized pipelines to support dynamic table extraction.



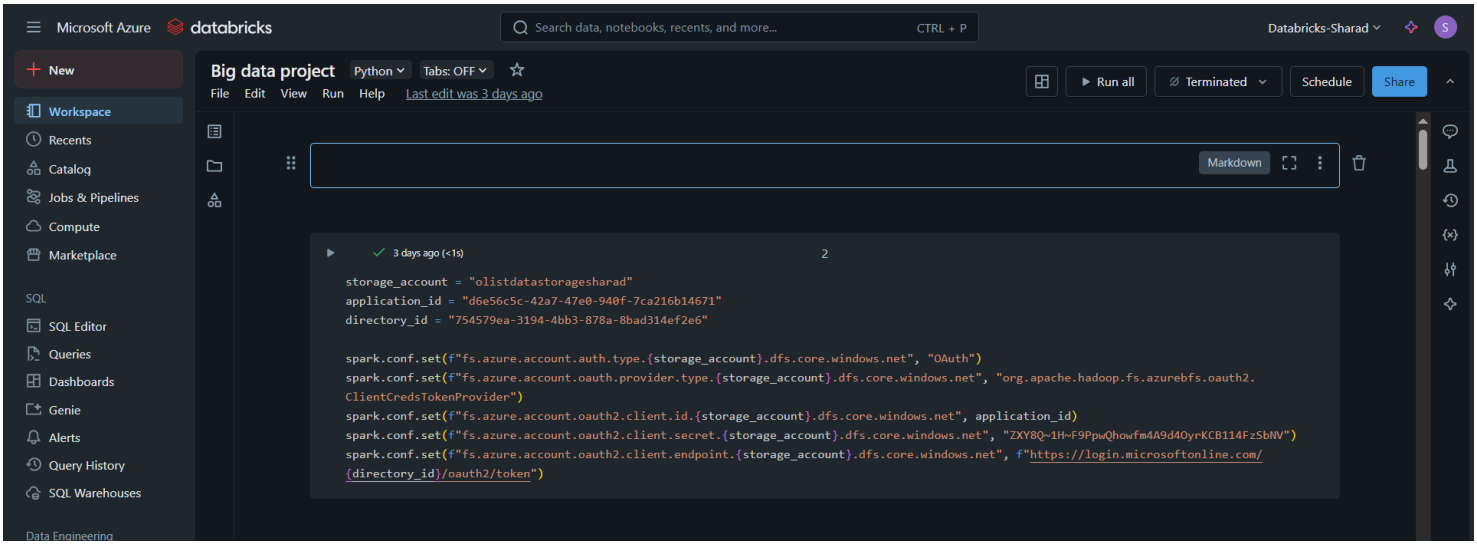
- Stored the raw data in the Bronze layer of ADLS (/bronze folder).



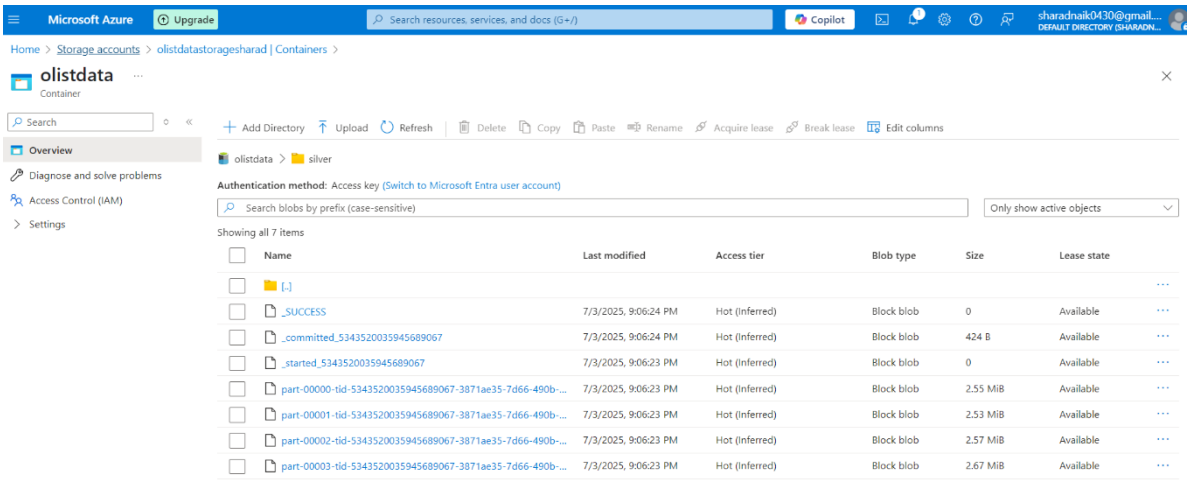
Name	Last modified	Access tier	Blob type	Size	Lease state
[..]					
olist_customers_dataset.csv	6/29/2025, 11:45:51 PM	Hot (Inferred)	Block blob	8.62 MiB	Available
olist_geolocation_dataset.csv	6/29/2025, 11:46:08 PM	Hot (Inferred)	Block blob	58.44 MiB	Available
olist_order_items_dataset.csv	6/29/2025, 11:46:23 PM	Hot (Inferred)	Block blob	14.72 MiB	Available
olist_order_payments_dataset.csv	6/29/2025, 11:47:45 PM	Hot (Inferred)	Block blob	6.28 MiB	Available
olist_order_reviews_dataset.csv	6/29/2025, 11:46:37 PM	Hot (Inferred)	Block blob	13.78 MiB	Available
olist_orders_dataset.csv	6/29/2025, 11:46:55 PM	Hot (Inferred)	Block blob	16.84 MiB	Available
olist_products_dataset.csv	6/29/2025, 11:47:11 PM	Hot (Inferred)	Block blob	2.27 MiB	Available
olist_sellers_dataset.csv	6/29/2025, 11:47:25 PM	Hot (Inferred)	Block blob	170.61 KiB	Available

Step 2: Data Cleaning & Transformation (Silver Layer)

- Connected Azure Databricks to ADLS Gen2 using service principal.



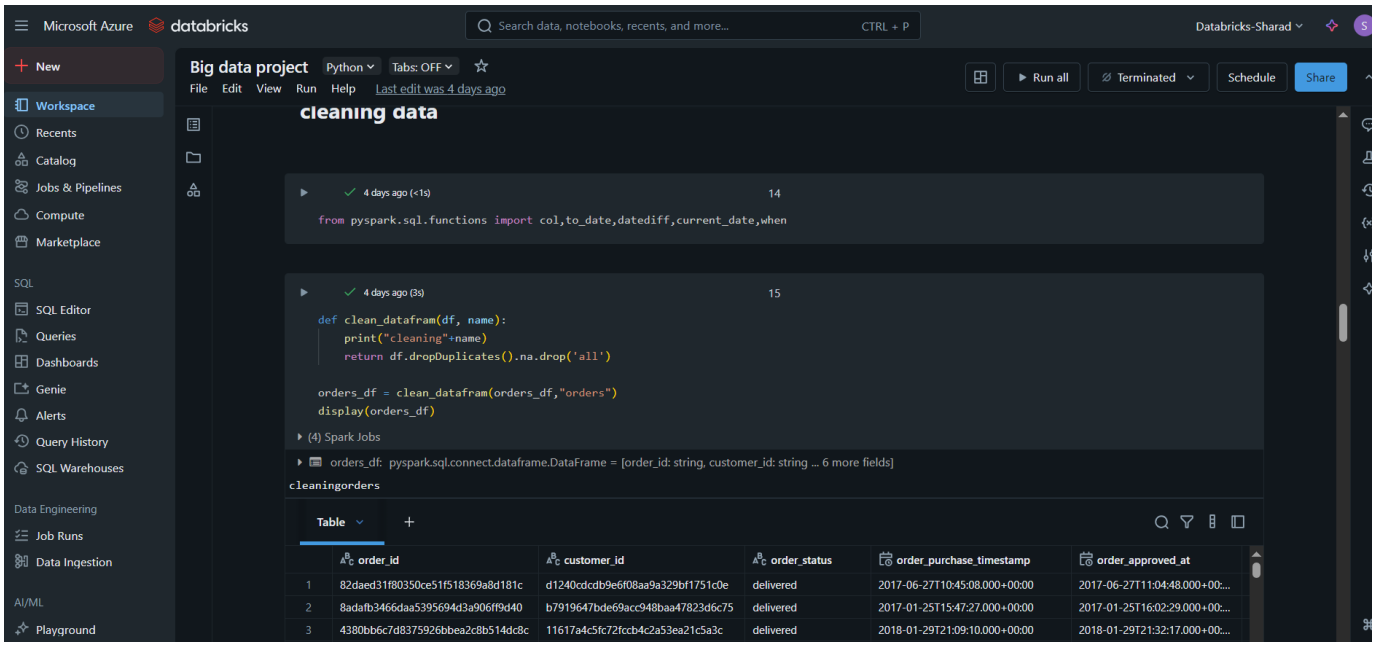
- Loaded data from the bronze folder for cleaning and transformation using PySpark.

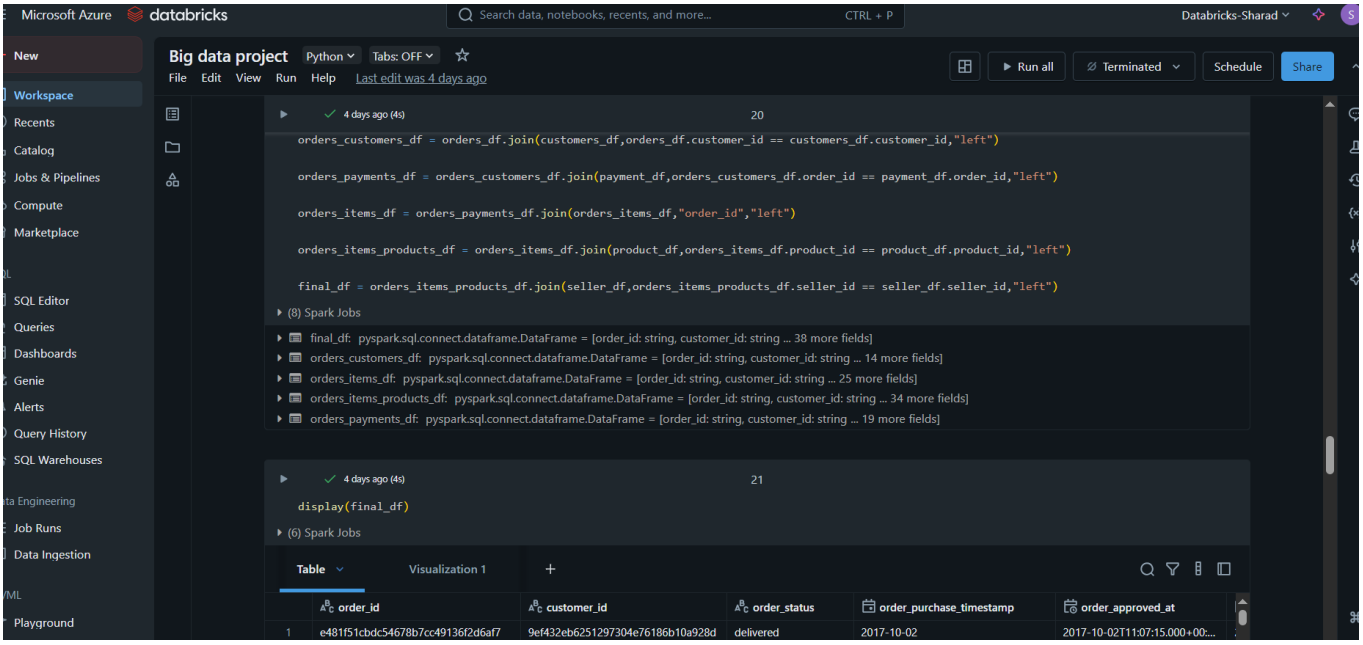


- Stored the processed data into the silver layer of ADLS (/silver folder).

Step 3: Enriching Data with MongoDB

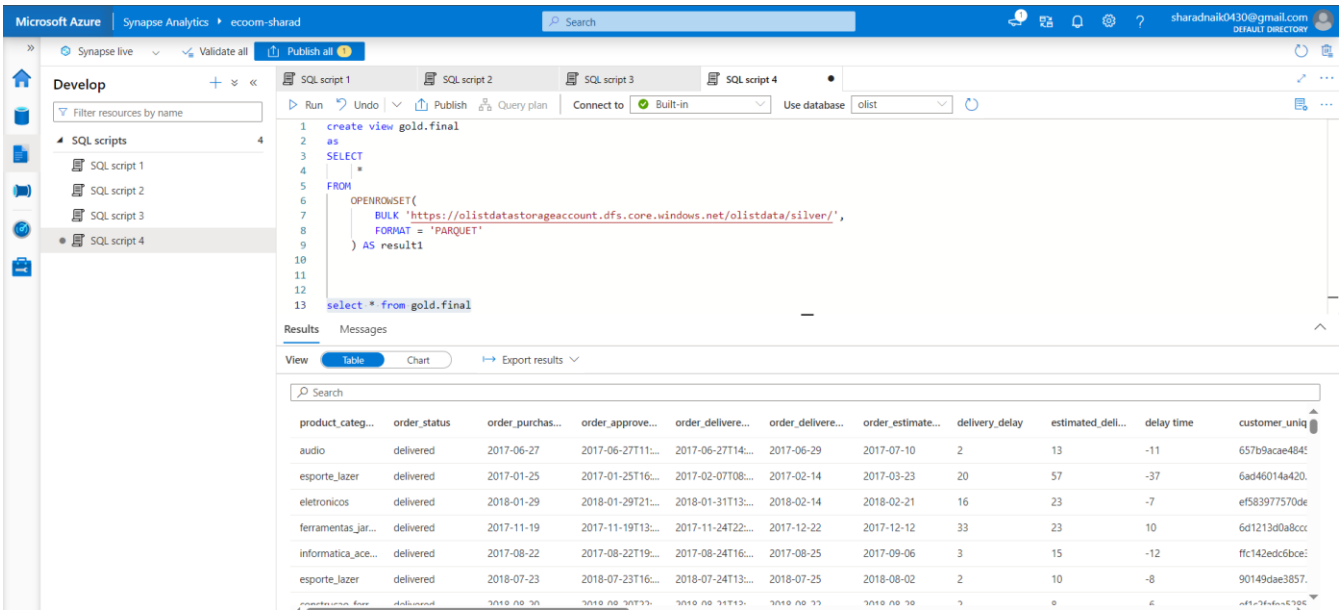
- Imported additional customer and product metadata from MongoDB into Databricks.
- Performed joins and enrichments with MySQL data.
- Saved the cleaned and enriched data into the silver folder.



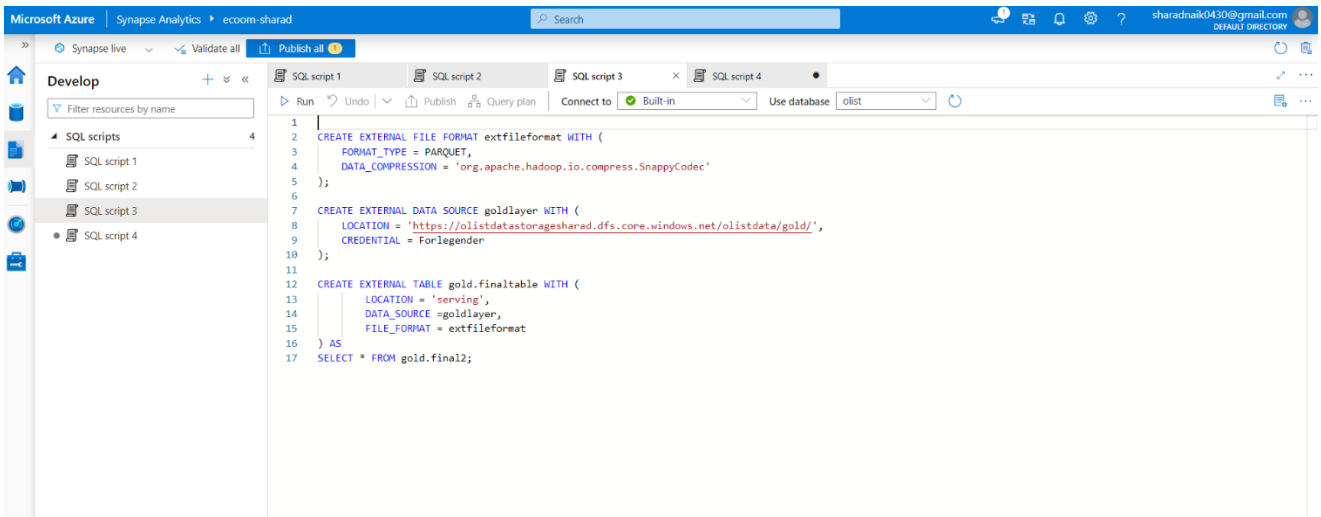


Step 4: Data Curation & Reporting (Gold Layer)

- Ingested the silver layer data into Azure Synapse Analytics for final curation.
- Created Synapse views and SQL scripts to generate metrics.
- Used CETAS to stored curated datasets back into the Gold layer in ADLS (/gold folder).



- CETAS script to store data to ADLS



Outcomes:

- Demonstrated a real-time batch data processing pipeline using industry-standard practices.
- Implemented secure and scalable connections between services.
- Achieved structured data layers that support reporting and analytics.