Stream Processing and Analysis

Taxi Data Analytics

Submitted by

S Yogesh Reddy - (2022BIFT07AED026)

In partial fulfilment for the award of the degree of

BACHELOR OF TECHNOLOGY

In

INFORMATION TECHNOLOGY

Under the Supervision of

Dr. Chetan J Shelke

DEPARTMENT OF INFORMATION TECHNOLOGY

ALLIANCE SCHOOL OF ADVANCED COMPUTING

ALLIANCE UNIVERSITY, BENGALURU

MARCH – 2025

Abstract

This project showcases the development of an end-to-end data engineering pipeline using Microsoft Azure. The objective is to design a streamlined data flow, from ingestion to visualization, enabling real-time insights into business data. By integrating services like Azure Data Factory, Azure Data Lake, Databricks, and Power BI, the pipeline automates data handling and transformation, delivering valuable business intelligence. The data used in this project is from Taxi, focusing on ride analytics to uncover patterns in demand, pricing, and customer behaviour.

INTRODUCTION

Modern businesses generate massive volumes of data, making it essential to have a robust data pipeline that can collect, transform, and visualize information. The ability to efficiently process and analyze data is critical for making informed business decisions and staying competitive. This project explores how Azure's cloud ecosystem can handle these tasks, transforming raw data into actionable insights that drive strategic decisions.

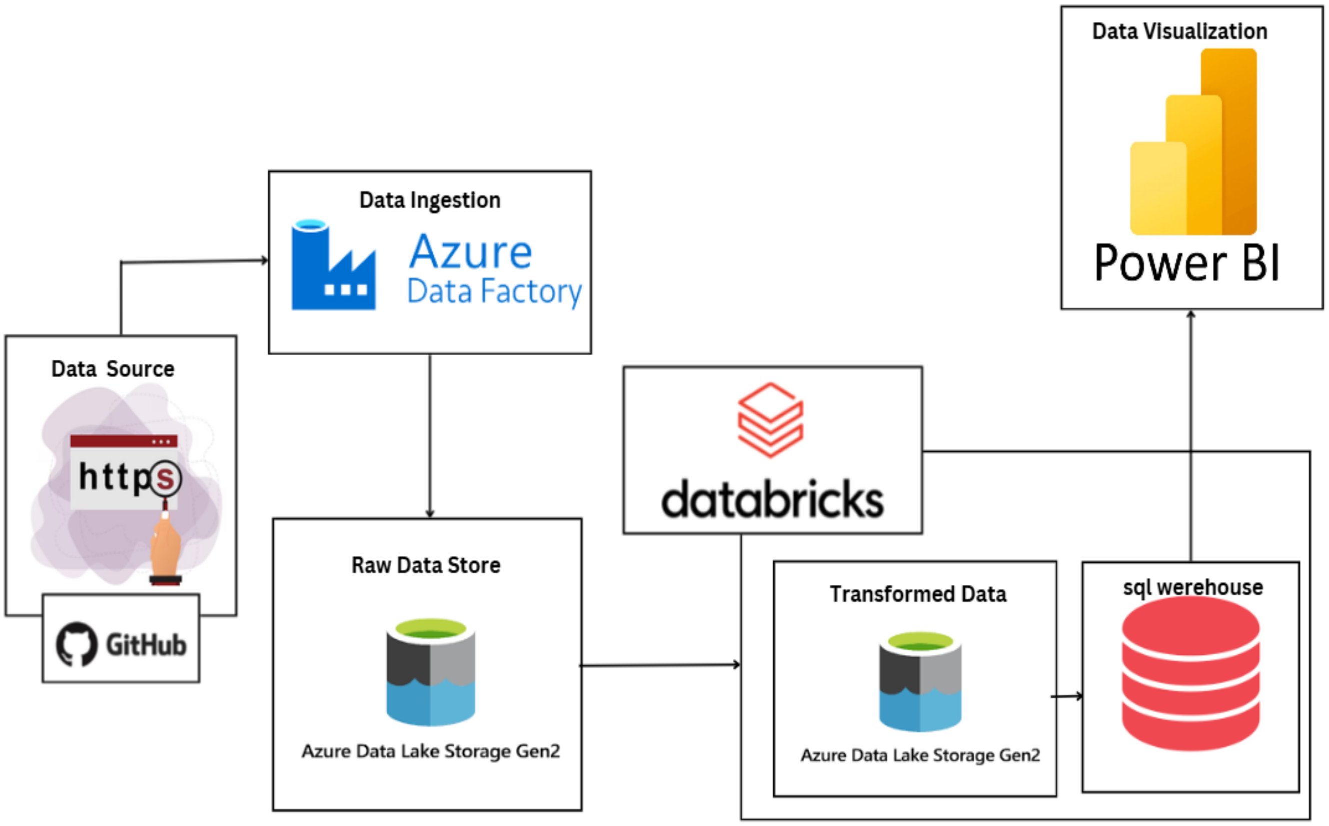
The data pipeline is designed to ingest Taxi ride data from GitHub, where ride records, fare amounts, customer ratings, pickup and drop-off locations, and timestamps are stored. Using Azure Data Factory, the data is automatically extracted and loaded into Azure Data Lake Storage Gen2. This raw data is then processed using Azure Databricks, where Spark-based transformations clean, aggregate, and enrich the data to prepare it for analysis.

Once transformed, the data is loaded into a SQL warehouse within Databricks, which acts as a central repository for structured analytical data. The SQL warehouse enables rapid querying and complex analytics, making it easy to derive valuable insights. Power BI is connected to this warehouse to create dynamic, interactive dashboards, allowing stakeholders to visualize trends like peak booking hours, revenue distribution across cities, and customer satisfaction metrics.

This end-to-end pipeline not only automates data ingestion and transformation but also ensures that reports are always up to date, giving business leaders access to the most current insights. By leveraging Azure's scalability and integration capabilities, the solution is well-equipped to handle increasing data volumes as the business grows, providing a future-proof foundation for data-driven decision-making.

METHODOLOGY

The Flowchart Diagram:

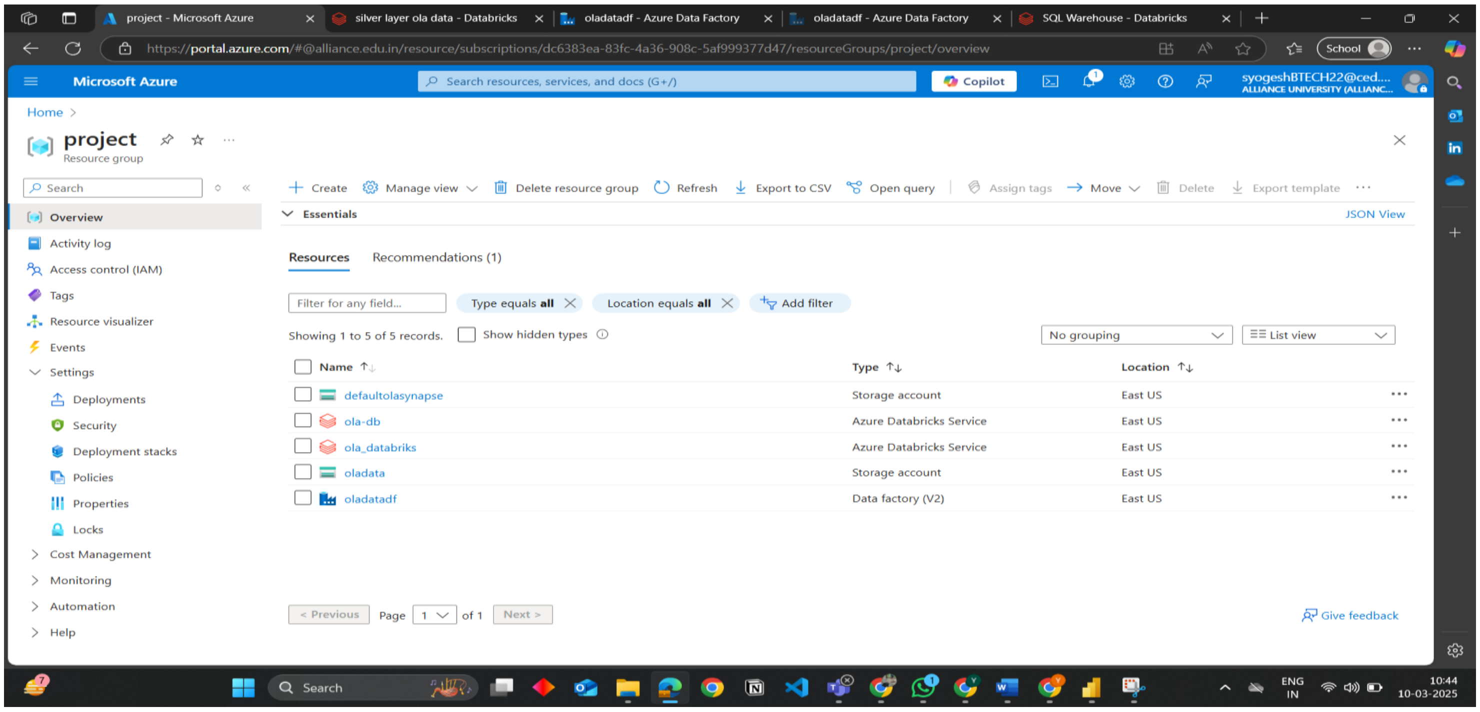


Step 1: Data Source

GitHub link for Data: https://github.com/Yogeshreddy07/OlaCabs-Data-Analytics-Azure.git

- The data originates from Taxi, accessed via GitHub over HTTPS.
- The dataset includes information about ride bookings, trip durations, pickup locations, drop-off points, fare amounts, and customer ratings.
- GitHub serves as a central repository for version-controlled data, making it a reliable source for real-time data ingestion.

Step 2: Resource Group:(Create the **-** Azure Data Factory, Storge Account, Databricks**)**



https://portal.azure.com/#@alliance.edu.in/resource/subscriptions/dc6383ea-83fc-4a36-908c-5af999377d47/resourceGroups/project/overview

Showing 1 to 5 of 5 records.

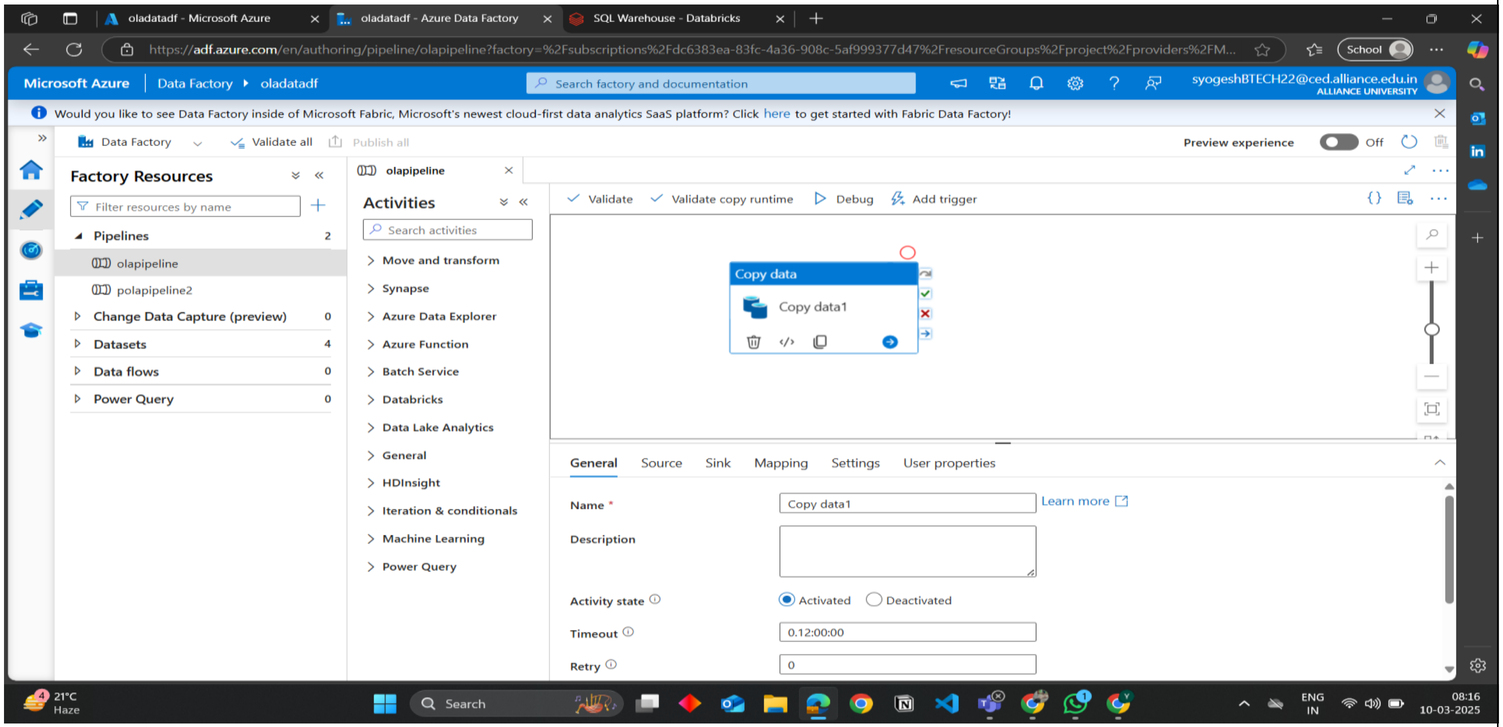
Name	Type	Location
defaultolasynapse	Storage account	East US
ola-db	Azure Databricks Service	East US
ola_databriks	Azure Databricks Service	East US
oladata	Storage account	East US
oladatadf	Data factory (V2)	East US

Step 3: Data Ingestion:

- Azure Data Factory (ADF) orchestrates the data ingestion process.
- ADF pipelines are set up to extract Taxi ride data and load it into Azure Data Lake Storage Gen2.

- Triggers and schedules are configured to automate the ingestion process, ensuring data freshness.

Pipeline Connection for Data Ingestion (Data Lake Storage Gen2**):**



https://adf.azure.com/en/authoring/pipeline/olapipeline?factory=%2Fsubscriptions%2Fdc6383ea-83fc-4a36-908c-5af999377d47%2FresourceGroups%2Fproject%2Fproviders%2FM...

Would you like to see Data Factory inside of Microsoft Fabric, Microsoft's newest cloud-first data analytics SaaS platform? Click here to get started with Fabric Data Factory!

Factory Resources	
Pipelines	2
olapipeline	
polapipeline2	
Change Data Capture (preview)	0
Datasets	4
Data flows	0
Power Query	0

Activities: Move and transform, Synapse, Azure Data Explorer, Azure Function, Batch Service, Databricks, Data Lake Analytics, General, HDInsight, Iteration & conditionals, Machine Learning, Power Query

Copy data — Copy data1

General | Source | Sink | Mapping | Settings | User properties

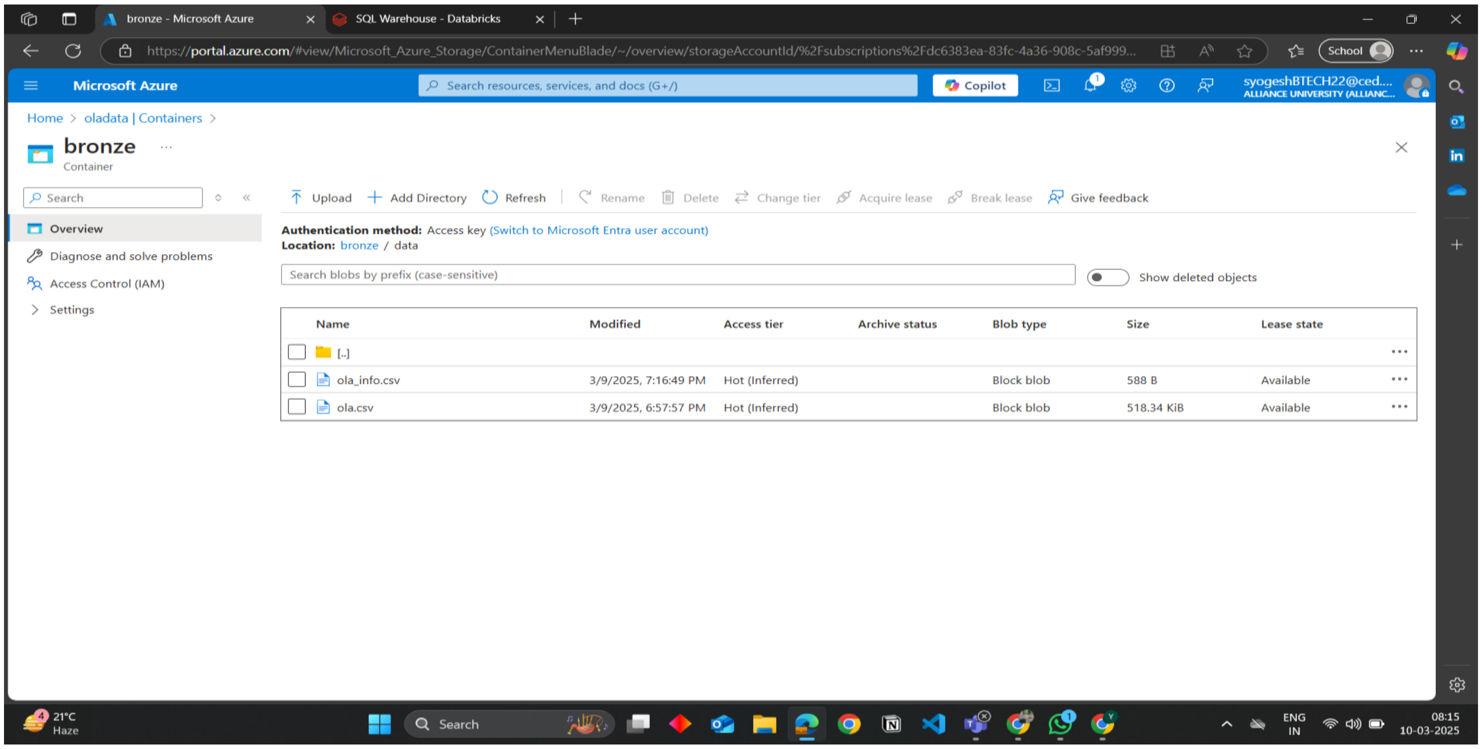
Name *	Copy data1
Description	
Activity state	[✓ Activated] Deactivated
Timeout	0.12:00:00
Retry	0

oladatadf - Microsoft Azure
oladatadf - Azure Data Factory
SQL Warehouse - Databricks
Microsoft Azure | Data Factory ▸ oladatadf
Search factory and documentation
syogeshBTECH22@ced.alliance.edu.in
ALLIANCE UNIVERSITY
Data Factory
Validate all
Publish all
Preview experience
Off
olapipeline
Validate
Validate copy runtime
Debug
Add trigger
Learn more
21°C
Haze
Search
ENG
IN
08:16
10-03-2025

Step 4: Raw Data Storage:

- Ingested data is stored in Azure Data Lake Storage Gen2.
- The lakehouse architecture is used, with raw data landing in the “bronze” layer.
- The data lake provides a scalable and cost-efficient storage solution, handling large volumes of ride and customer data.

Data loaded into Storage account- Containers:


bronze - Microsoft Azure
SQL Warehouse - Databricks
https://portal.azure.com/#view/Microsoft_Azure_Storage/ContainerMenuBlade/~/overview/storageAccountId/%2Fsubscriptions%2Fdc6383ea-83fc-4a36-908c-5af999...
Microsoft Azure
Search resources, services, and docs (G+/)
Copilot
syogeshBTECH22@ced....
ALLIANCE UNIVERSITY (ALLIANC...
Home > oladata | Containers >
bronze
Container
Search
Overview
Diagnose and solve problems
Access Control (IAM)
Settings
Upload
Add Directory
Refresh
Rename
Delete
Change tier
Acquire lease
Break lease
Give feedback

Authentication method: Access key (Switch to Microsoft Entra user account)
Location: bronze / data

Search blobs by prefix (case-sensitive)   Show deleted objects

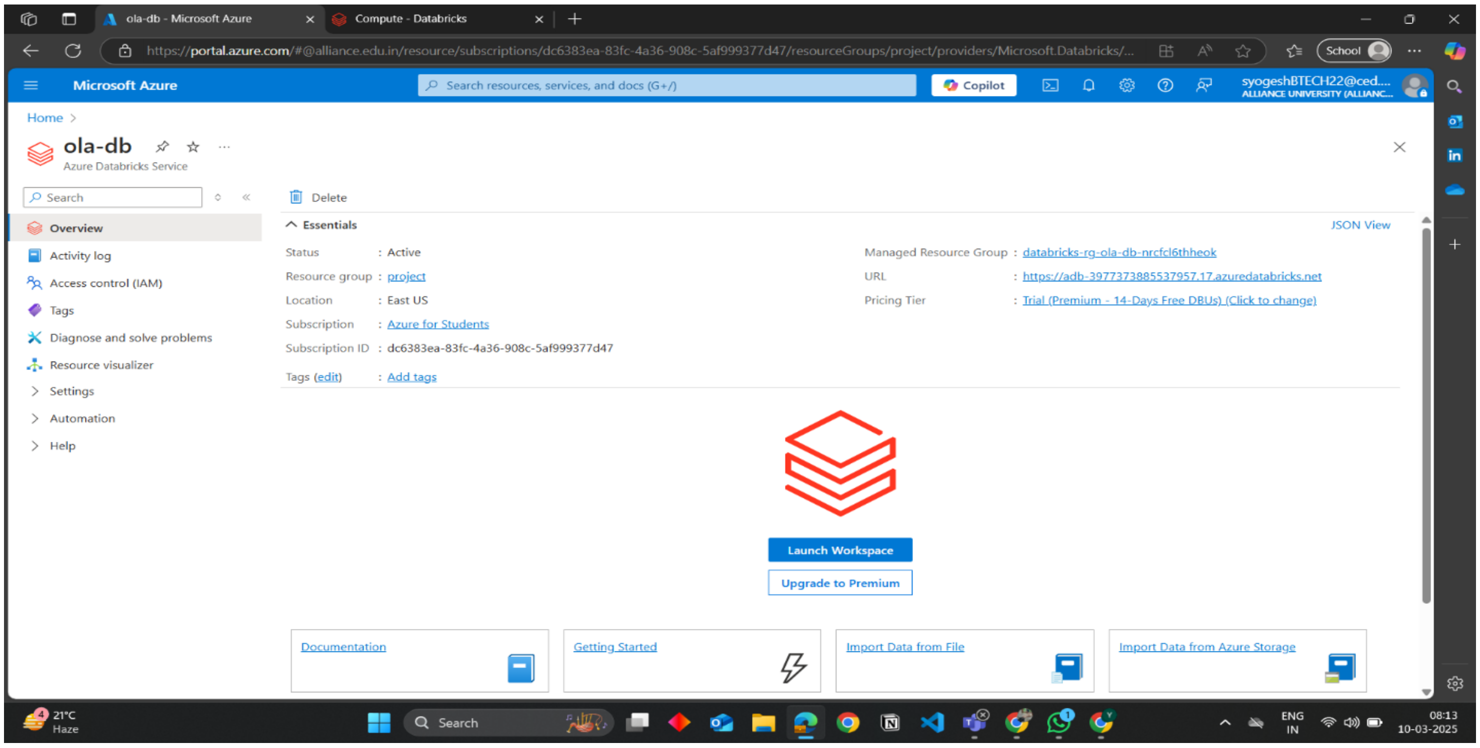
Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
[..]						
ola_info.csv	3/9/2025, 7:16:49 PM	Hot (Inferred)		Block blob	588 B	Available
ola.csv	3/9/2025, 6:57:57 PM	Hot (Inferred)		Block blob	518.34 KiB	Available

21°C
Haze
Search
ENG
IN
08:15
10-03-2025

Step 5: Data Transformation:

- Azure Databricks, powered by Apache Spark, processes the raw data.
- Notebooks are created in Databricks to clean, enrich, and aggregate the ride data.
- The transformation process includes handling missing values, calculating ride distances, aggregating revenue by location, and analyzing peak hours.
- The data is organized into bronze (raw), silver (cleaned), and gold (aggregated) layers, following the medallion architecture.

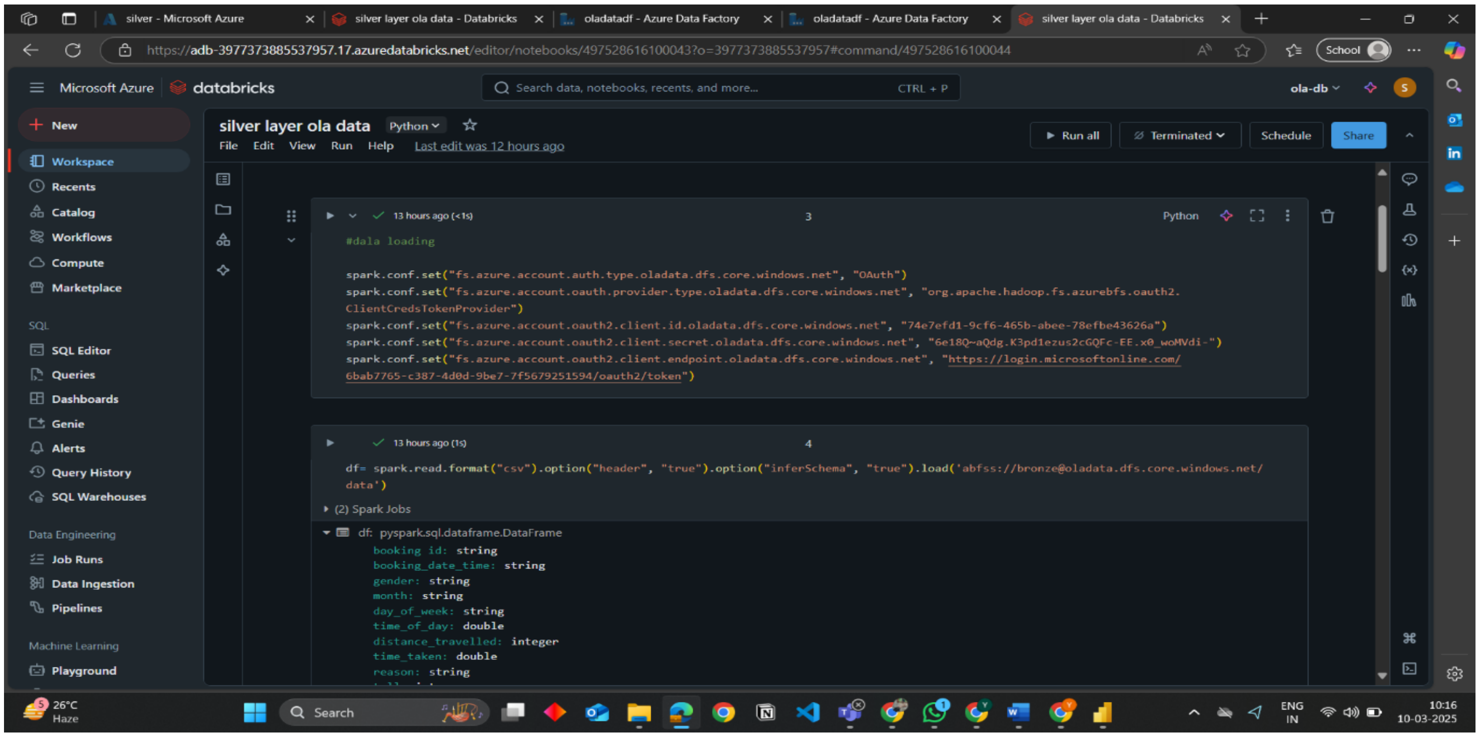
Creation of Databricks:



ola-db
Azure Databricks Service

Status	: Active
Resource group	: project
Location	: East US
Subscription	: Azure for Students
Subscription ID	: dc6383ea-83fc-4a36-908c-5af999377d47
Tags (edit)	: Add tags
Managed Resource Group	: databricks-rg-ola-db-nrcfcl6thheok
URL	: https://adb-3977373885537957.17.azuredatabricks.net
Pricing Tier	: Trial (Premium - 14-Days Free DBUs) (Click to change)

Databricks Spark integration for loading Transformed data to silver:



silver layer ola data

```
#dala loading

spark.conf.set("fs.azure.account.auth.type.oladata.dfs.core.windows.net", "OAuth")
spark.conf.set("fs.azure.account.oauth.provider.type.oladata.dfs.core.windows.net", "org.apache.hadoop.fs.azurebfs.oauth2.
ClientCredsTokenProvider")
spark.conf.set("fs.azure.account.oauth2.client.id.oladata.dfs.core.windows.net", "74e7efd1-9cf6-465b-abee-78efbe43626a")
spark.conf.set("fs.azure.account.oauth2.client.secret.oladata.dfs.core.windows.net", "6e18Q~aQdg.K3pd1ezus2cGQFc-EE.x0_woMVdi-")
spark.conf.set("fs.azure.account.oauth2.client.endpoint.oladata.dfs.core.windows.net", "https://login.microsoftonline.com/
6bab7765-c387-4d0d-9be7-7f5679251594/oauth2/token")
```

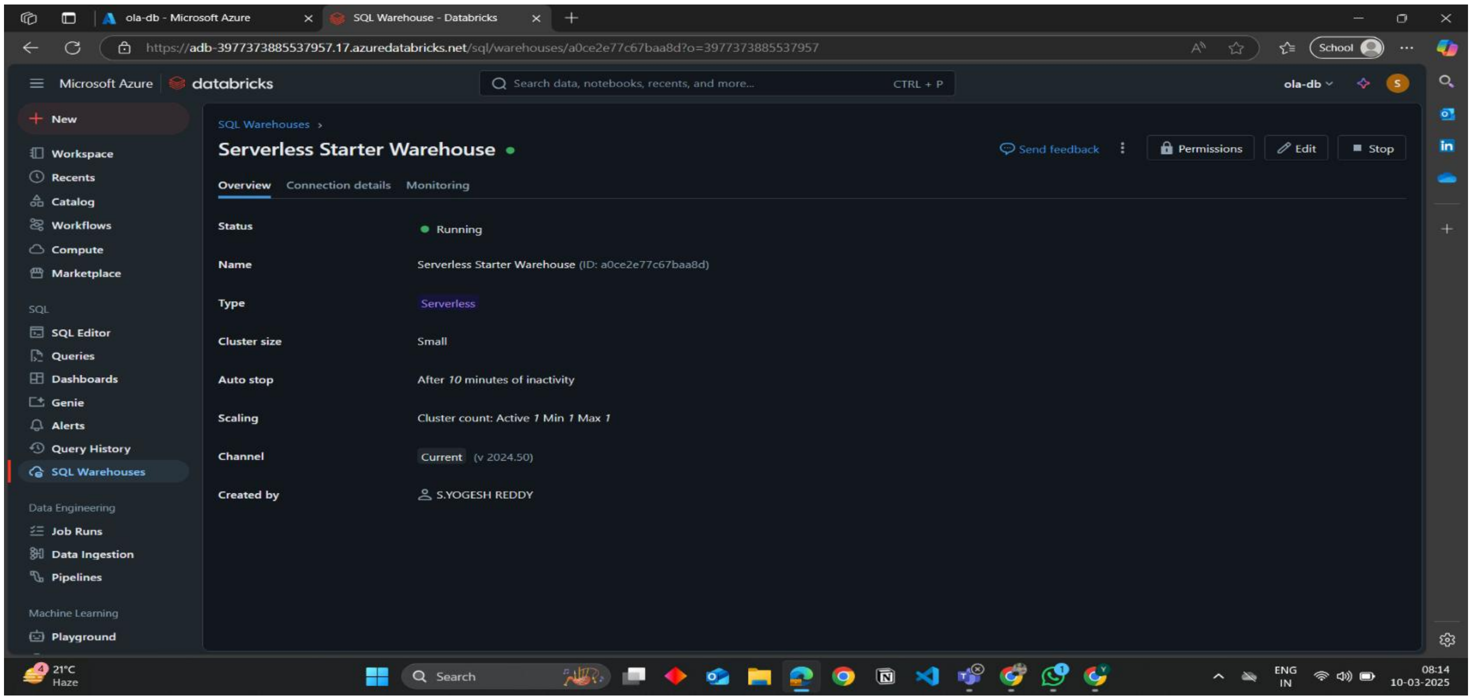
```
df= spark.read.format("csv").option("header", "true").option("inferSchema", "true").load('abfss://bronze@oladata.dfs.core.windows.net/
data')
```

```
(2) Spark Jobs
df: pyspark.sql.dataframe.DataFrame
    booking_id: string
    booking_date_time: string
    gender: string
    month: string
    day_of_week: string
    time_of_day: double
    distance_travelled: integer
    time_taken: double
    reason: string
```

SQL Warehouse:

- Transformed data is loaded into a SQL warehouse within Databricks.
- The SQL warehouse enables fast, efficient querying, supporting complex analytical workloads.
- It acts as the central repository for analytical data, ready to be consumed by business intelligence tools.

SQL Warehouse Server Integration



SQL Warehouses >

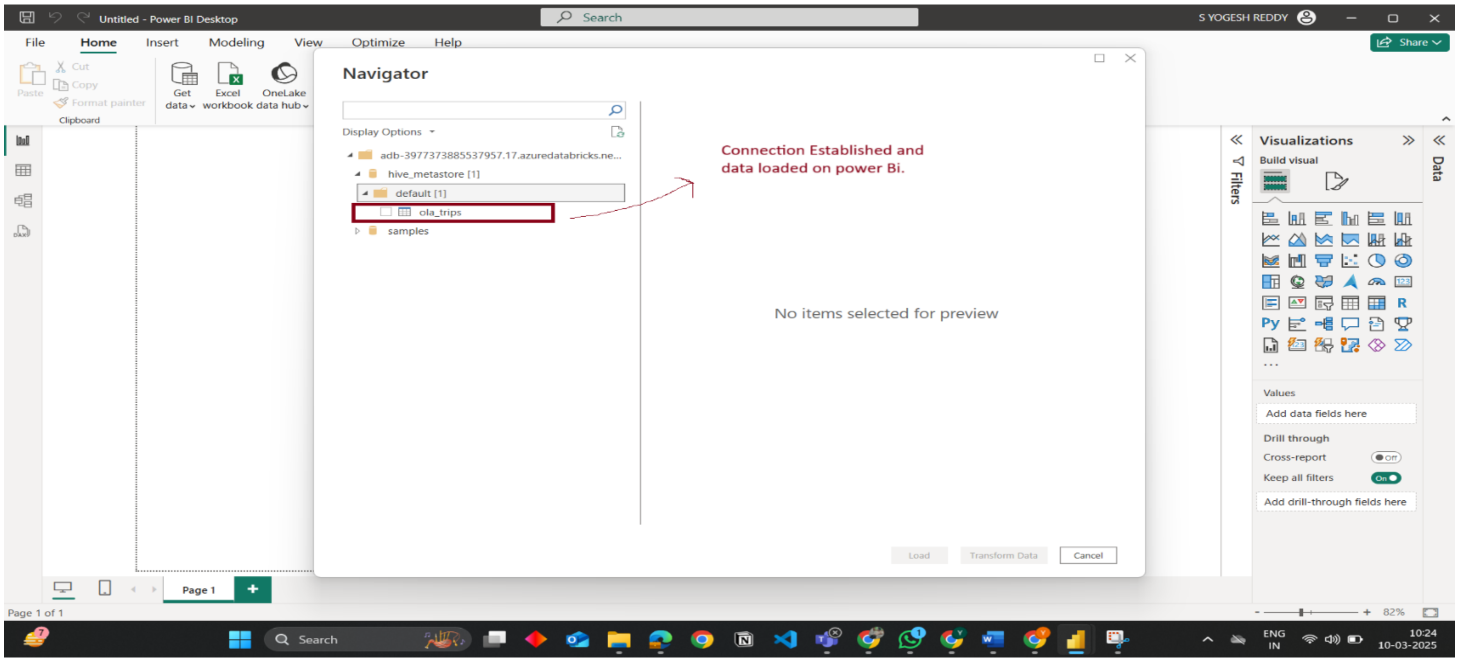
Serverless Starter Warehouse

Status	Running
Name	Serverless Starter Warehouse (ID: a0ce2e77c67baa8d)
Type	Serverless
Cluster size	Small
Auto stop	After 10 minutes of inactivity
Scaling	Cluster count: Active 1 Min 1 Max 1
Channel	Current (v 2024.50)
Created by	S.YOGESH REDDY

Data Visualization:

- Power BI connects to the SQL warehouse to build interactive dashboards.
- Visualizations include ride volume trends, revenue by city, average fare amounts, customer ratings, and trip heatmaps.
- The dashboard provides filtering capabilities, allowing users to explore data by date, location, fare range, and customer feedback.

Data integrated to Power BI :

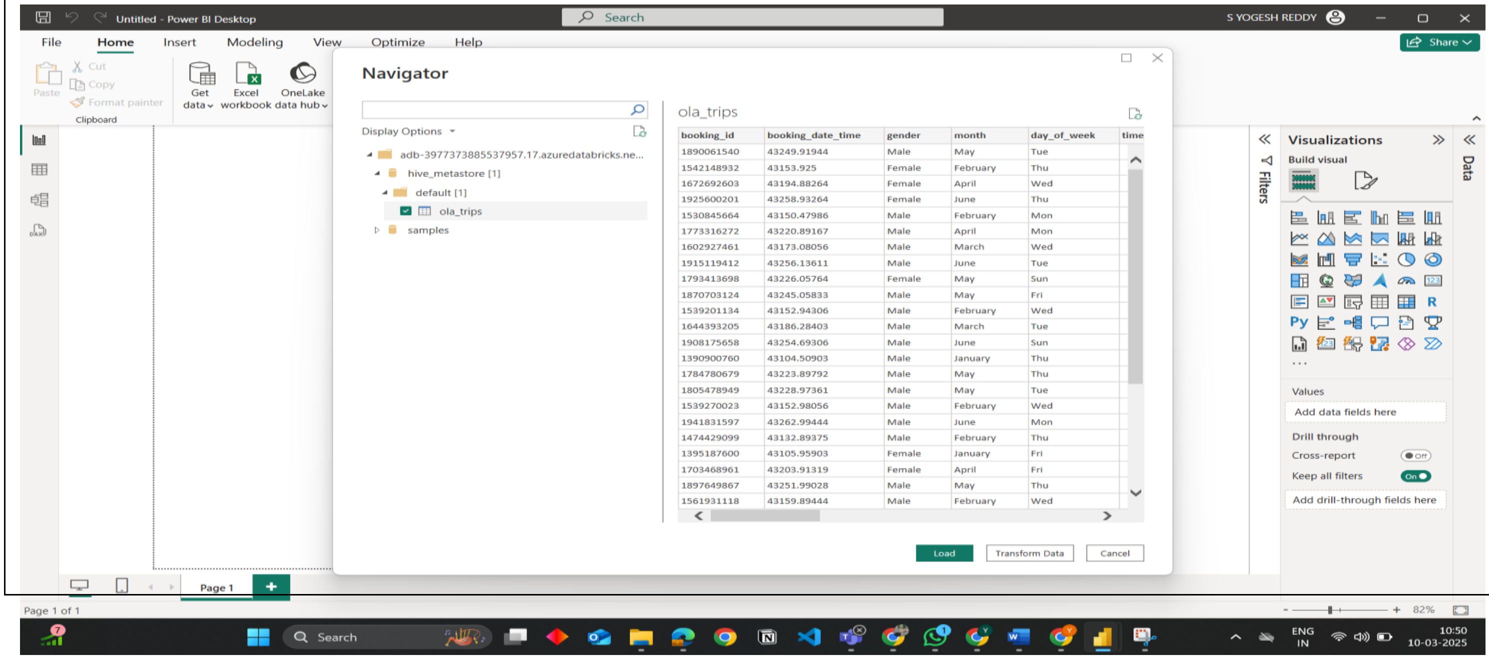


Connection Established and data loaded on power Bi.

RESULTS

The pipeline delivers a fully automated solution, enabling real-time data analysis for Ola Cabs. Stakeholders can access up-to-date dashboards that provide a clear view of operational metrics. The SQL warehouse ensures rapid query execution, while Power BI offers dynamic, easy-to-understand visualizations. The solution is scalable, allowing future growth in data volume and complexity.

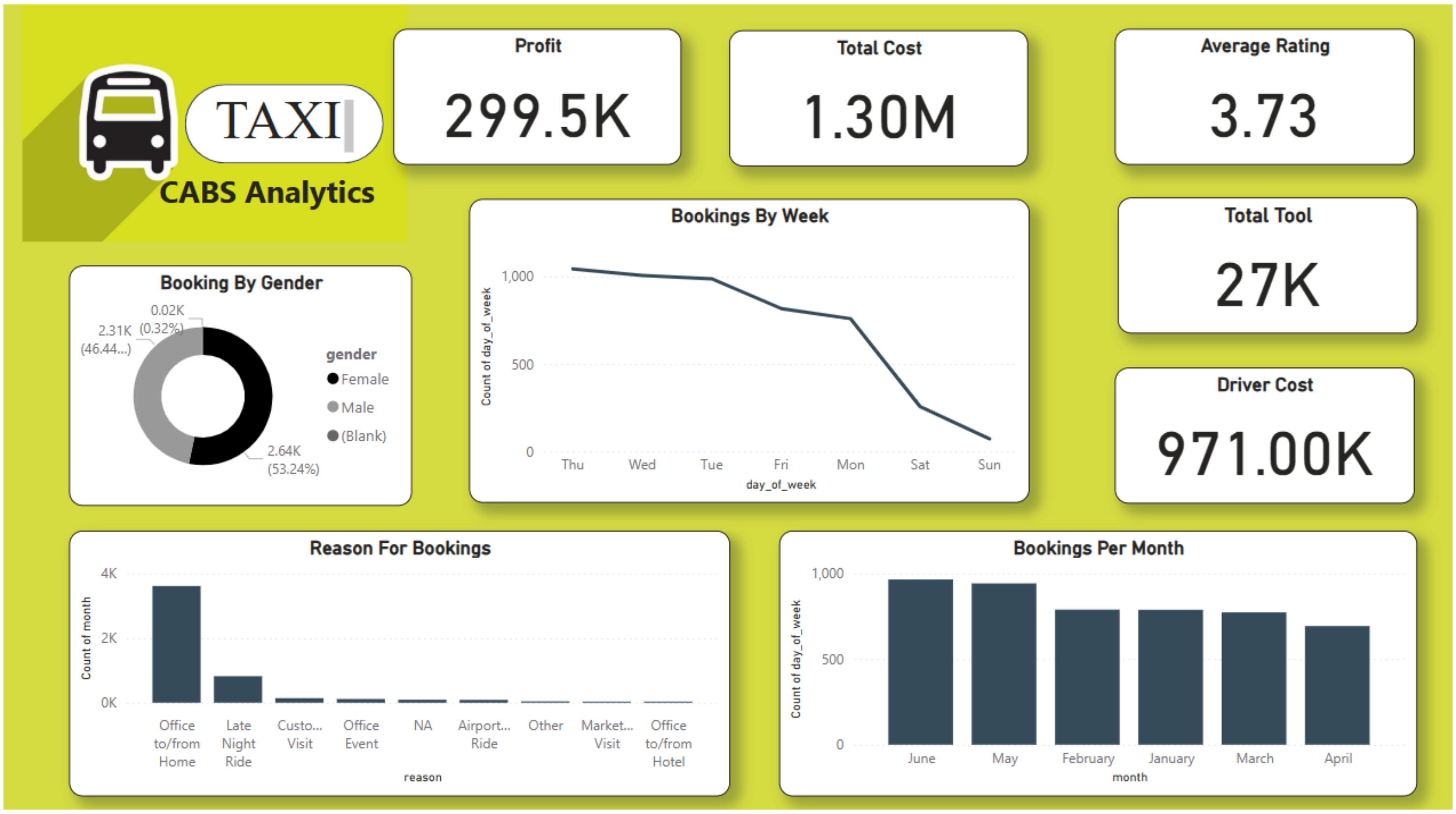
Data Loaded in to Power BI:



ola_trips

booking_id	booking_date_time	gender	month	day_of_week	time
1890061540	43249.91944	Male	May	Tue	
1542148932	43153.925	Female	February	Thu	
1672692603	43194.88264	Female	April	Wed	
1925600201	43258.93264	Female	June	Thu	
1530845664	43150.47986	Male	February	Mon	
1773316272	43220.89167	Male	April	Mon	
1602927461	43173.08056	Male	March	Wed	
1915119412	43256.13611	Male	June	Tue	
1793413698	43226.05764	Female	May	Sun	
1870703124	43245.05833	Male	May	Fri	
1539201134	43152.94306	Male	February	Wed	
1644393205	43186.28403	Male	March	Tue	
1908175658	43254.69306	Male	June	Sun	
1390900760	43104.50903	Male	January	Thu	
1784780679	43223.89792	Male	May	Thu	
1805478949	43228.97361	Male	May	Tue	
1539270023	43152.98056	Male	February	Wed	
1941831597	43262.99444	Male	June	Mon	
1474429099	43132.89375	Male	February	Thu	
1395187600	43105.95903	Female	January	Fri	
1703468961	43203.91319	Female	April	Fri	
1897649867	43251.99028	Male	May	Thu	
1561931118	43159.89444	Male	February	Wed	

The Power BI Output:



Conclusion

This project highlights the capabilities of Azure's integrated services in building a scalable, efficient, and fully automated data pipeline. By leveraging tools like Azure Data Factory, Databricks, and Power BI, the solution transforms raw Ola Cabs data into valuable business insights. It serves as an excellent learning experience, reinforcing core data engineering concepts and best practices in cloud architecture.