Project-2 (Data Engineering Batch-1)
Data Analysis using SparkSQL

Vuthur Sriganga

Hexaware Data Engineering Batch-1

vuthursriganga@gmail.com

Project-2 (Data Engineering Batch-1)

Data Analysis using SparkSQL

Title of Project:

Implement Data analysis using Spark SQL on Azure Databricks & Process data for errors , Seasonality, and Anomalies.

Project Overview:

The goal of this project is to implement data analysis using Spark SQL on Azure Databricks and process the data for errors, seasonality, and anomalies. The project involves leveraging Azure Databricks for its Spark-based analytics capabilities and Spark SQL for querying and processing large datasets efficiently.

About the project:

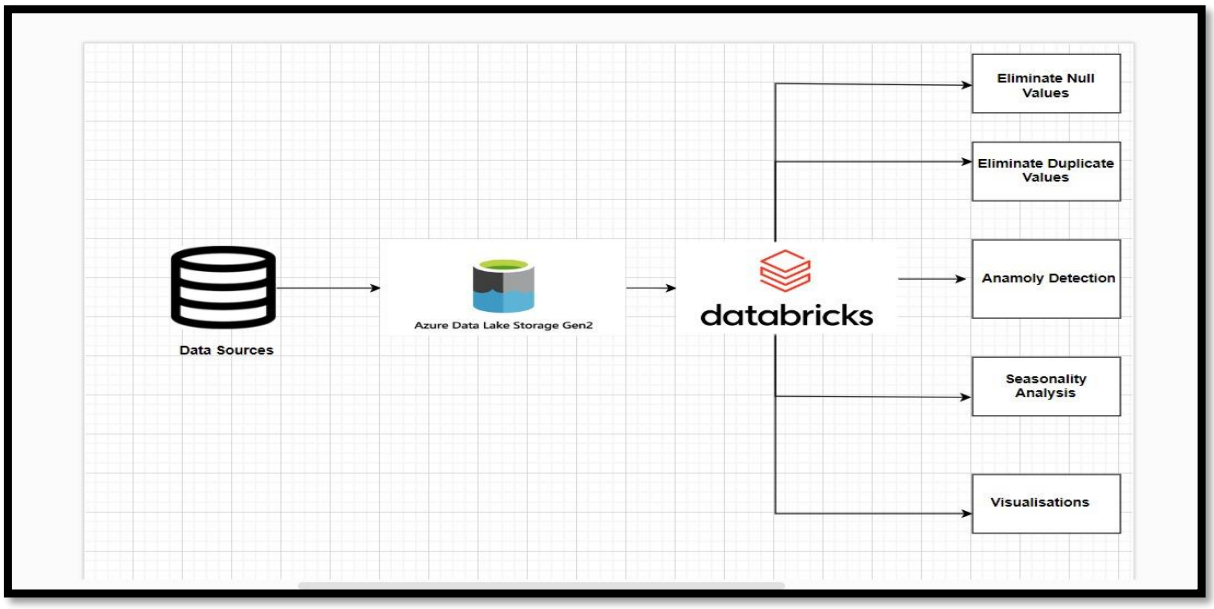
The project involves implementing data analysis using Spark SQL on Azure Databricks and processing the data for errors, seasonality, and anomalies. This initiative aims to leverage the scalability and performance of Spark on the Azure Databricks platform for efficient data processing and analysis.

Key Components of this project:

1) Data Ingestion
2) Data Exploration
3) Data Cleaning
4) Data Transformation
5) Data Analysis with Spark SQL
6) Error Detection
7) Seasonality Analysis
8) Anomaly Detection
9) Visualization and Reporting

Architecture Diagram:

The Architecture diagram shows the flow of execution of project and the operations that are performed on the dataset.



How it works

1) Setup Azure Databricks Environment:

- Create an Azure Databricks workspace in the Azure portal.
- Choose the appropriate subscription, resource group, and workspace name.
- Configure the pricing tier based on your requirements. Options include Standard and Premium tiers with different capabilities and performance levels.
- Once the deployment is complete, click on Go to Resource.
- Create a new notebook in the workspace.
- Create a compute by providing the required details and configure it with data bricks notebook.

2) Data Ingestion:

- Create an Azure Data Lake Storage Gen2 account in Azure Portal.
- Enable the Hierarchical Namespace in the Advanced Settings of ADLS account.
- Create a container in the ADLS Gen2 account.Upload the source file into container.
- Configure the Azure Data Lake Storage Gen2 account with Data Bricks to perform all the required operations.

3) Data Exploration and Cleaning:

- Explore the ingested data using Databricks notebooks, which provide an interactive environment for data exploration and analysis.

- Use Spark DataFrame APIs or Spark SQL within the notebooks to examine the data's schema, sample rows.
- Identify data quality issues such as missing values, duplicate records, inconsistent formats, and outliers.
- Apply data cleaning techniques such as removing null values, deduplication to ensure data quality and consistency.

4) Data Transformation:

- Transform the cleaned data into a format suitable for analysis using Spark SQL.
- Perform transformations such as filtering, aggregating to prepare the data for analysis.
- Leverage Spark's distributed processing capabilities to handle large-scale data transformations efficiently.

5) Data Analysis with Spark SQL:

- Write Spark SQL queries within Databricks notebooks to perform advanced analytics on the transformed data.
- Utilize SQL syntax to extract insights, perform aggregations, and apply statistical functions.
- Leverage Spark SQL's built-in functions and windowing functions for complex analytics tasks such as time series analysis, Seasonality detection, and anomaly detection.

6) Error Detection:

- Implement algorithms or techniques within Spark SQL to detect errors or anomalies in the data.
- Utilize SQL queries and functions to identify inconsistencies or patterns deviating from expected behavior.

7) Seasonality Analysis:

- Analyze the data for seasonality patterns using Spark SQL.
- Apply time series analysis techniques such as moving averages, or Fourier transforms to identify recurring trends or fluctuations over time.
- Use SQL queries and functions to aggregate data over time intervals and calculate seasonal metrics such as seasonal averages.

Steps:

- Apply moving average to smooth out seasonal variations.
- Analyze Moving Averages
- Visualize Moving Averages in Azure Data Bricks.

8) Anomaly Detection:

- Develop models or algorithms within Spark SQL to detect anomalies in the data.
- Utilize machine learning algorithms, statistical methods, or rule-based approaches for anomaly detection.

Steps:

- Calculate Z-scores for each value
- Define anomaly threshold (e.g., Z-score greater than 1)
- Analyze the identified anomalies
- Visualization of Anomalies (e.g., scatter plot of ORDERDATE vs. QUANTITYORDERED)

9) Visualization and Reporting:

- Visualize the analyzed data and insights using Databricks notebooks or external visualization tools.
- Use visualization in Databricks' built-in visualization tools to create compelling visualizations that highlight key insights and trends.

Azure Resources / Tools used for this Project:

1) Azure Data Lake Storage Gen2 (ADLS Gen2):

- ADLS Gen2 is the storage solution where the Olympics data will be ingested.
- ADLS Gen2 provides a scalable and cost-effective storage solution for big data analytics workloads.

2) Azure Databricks:

- Azure Databricks is a powerful analytics platform that enables organizations to use the capabilities of Apache Spark for processing big data in a collaborative and scalable environment.
- PySpark notebooks are be used in Azure Databricks to perform data transformation tasks like filtering ,group by and aggregations, selecting distinct data, renaming columns etc., on the ingested Olympics data.

3) Spark SQL:

- Spark SQL is a component of Apache Spark that introduces a structured data processing model along with a SQL-like query language.
- It allows users to run SQL queries programmatically and seamlessly integrate SQL queries with existing Spark programs.

4) Azure Portal:

The Azure Portal is a web-based interface for managing Azure services. It would be used for provisioning and configuring resources such as Azure Data Lake Storage Gen2, Azure Data Factory, and Azure Databricks.

5) Microsoft Excel :

Microsoft Excel might be used for data preparation or as a data source for creating input datasets for the project.

Project Requirements:

1) Data Ingestion:

Ingest raw data from various sources such as databases, files, or streaming sources into Azure Databricks storage.

2) Data Exploration:

Explore the ingested data to understand its structure, schema, and quality.

3) Data Cleaning:

- Clean the data to handle inconsistencies, missing values, and outliers.
- Ensure data consistency and reliability for further analysis.

4) Data Transformation:

- Transform the cleaned data into a suitable format for analysis.
- Reshape, aggregate, or derive new features as needed for analysis.

5) Data Analysis with Spark SQL:

- Utilize Spark SQL to perform advanced analytics on the transformed data.
- Write SQL queries to extract insights, perform aggregations, and apply statistical functions.

6) Error Detection:

- Implement algorithms or techniques to detect errors in the data.
- Identify inconsistencies, anomalies, or patterns deviating from expected behavior.

7) Seasonality Analysis:

Analyze the data for seasonality patterns to understand recurring trends or fluctuations over time.

8) Anomaly Detection:

- Develop models or algorithms to detect anomalies in the data.
- Use outlier detection, anomaly scoring (using z-score) , or machine learning-based techniques for anomaly detection.

9) Visualization and Reporting:

- Visualize analyzed data and insights using interactive visualizations or reports.
- Communicate findings to stakeholders for data-driven decision-making.

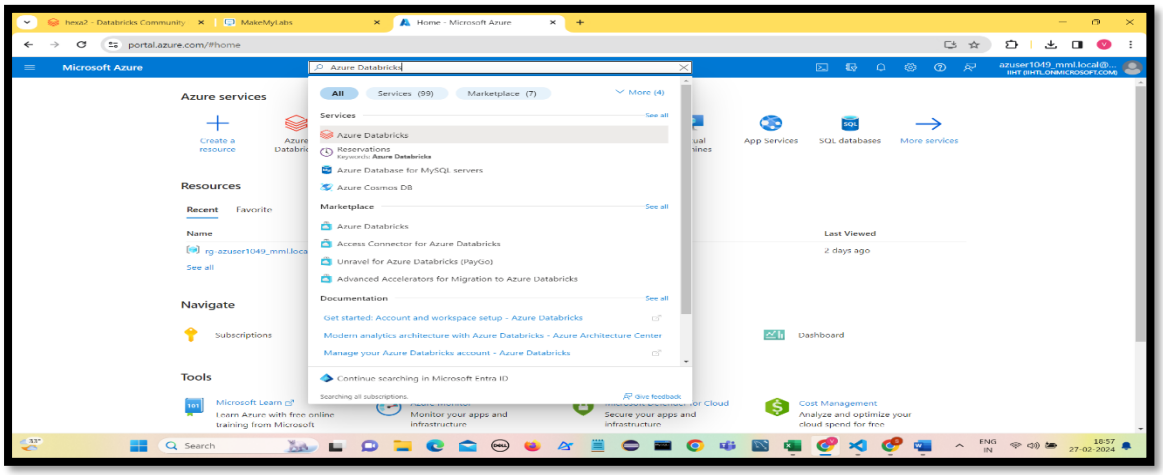
Overall Execution Flow:

1) Provision Azure Databricks workspace in Azure portal.
2) Ingest raw data from sources into Azure Databricks.
3) Explore data using Databricks notebooks for structure and quality.
4) Clean data by handling missing values, duplicates, and inconsistencies.
5) Transform cleaned data into suitable format for analysis.
6) Apply transformations such as reshaping and aggregating using Spark SQL.
7) Write Spark SQL queries in Databricks notebooks for advanced analytics.
8) Analyze data for errors, seasonality patterns, and anomalies using SQL queries.
9) Implement algorithms in Spark SQL for error detection and anomaly detection.
10) Analyze data for seasonality patterns using SQL queries and functions.

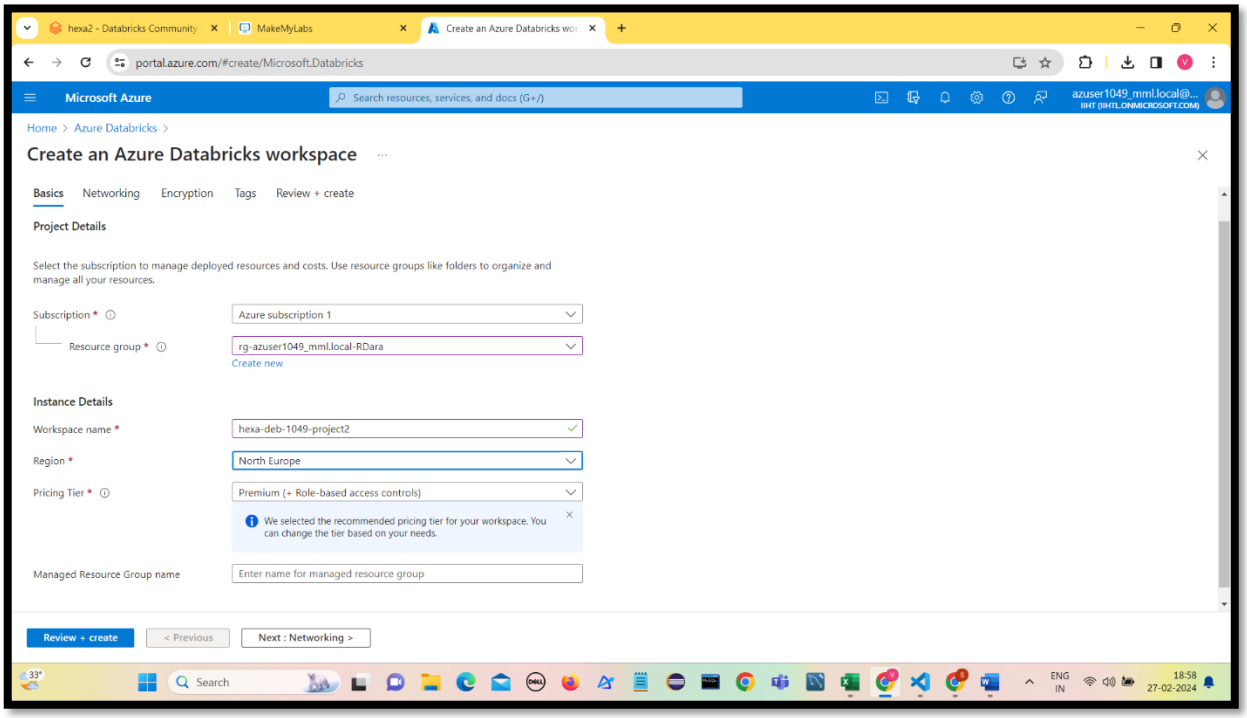
Execution of Project:

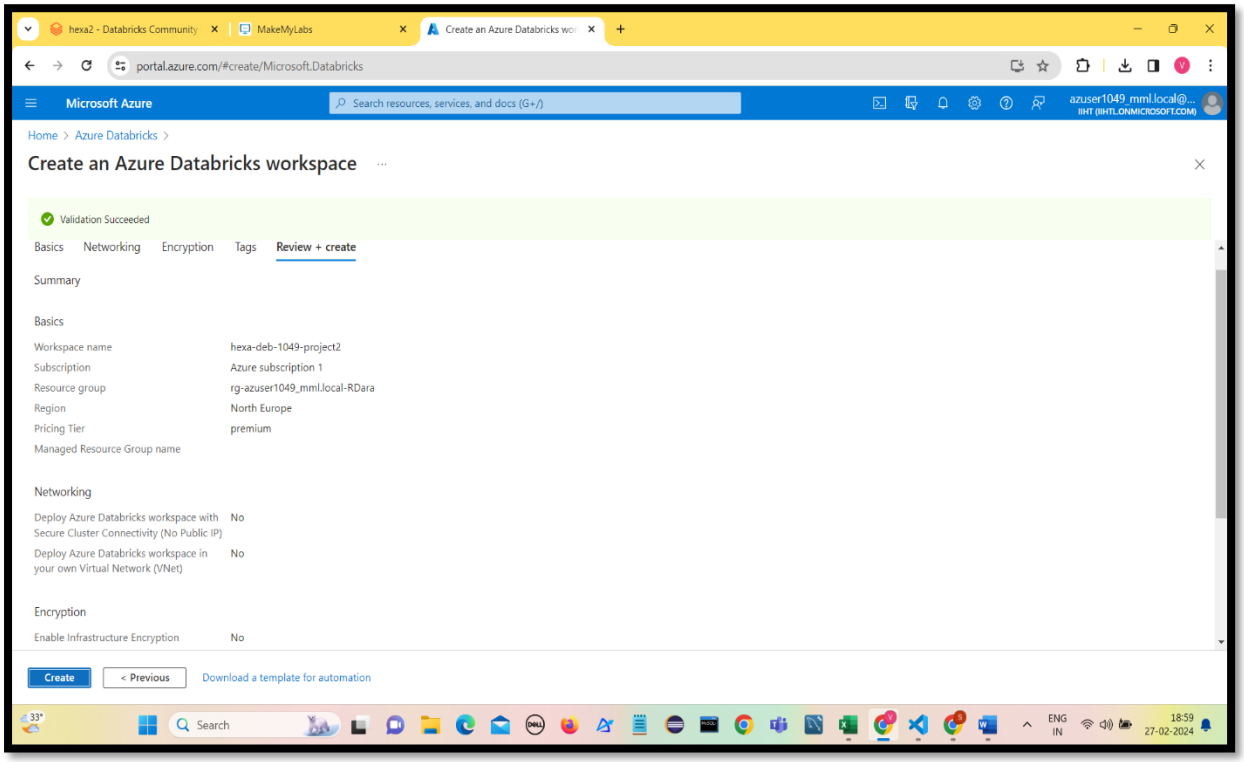
Creating Azure Databricks Workspace

1) Log in into the Azure Portal and Search for Azure Databricks

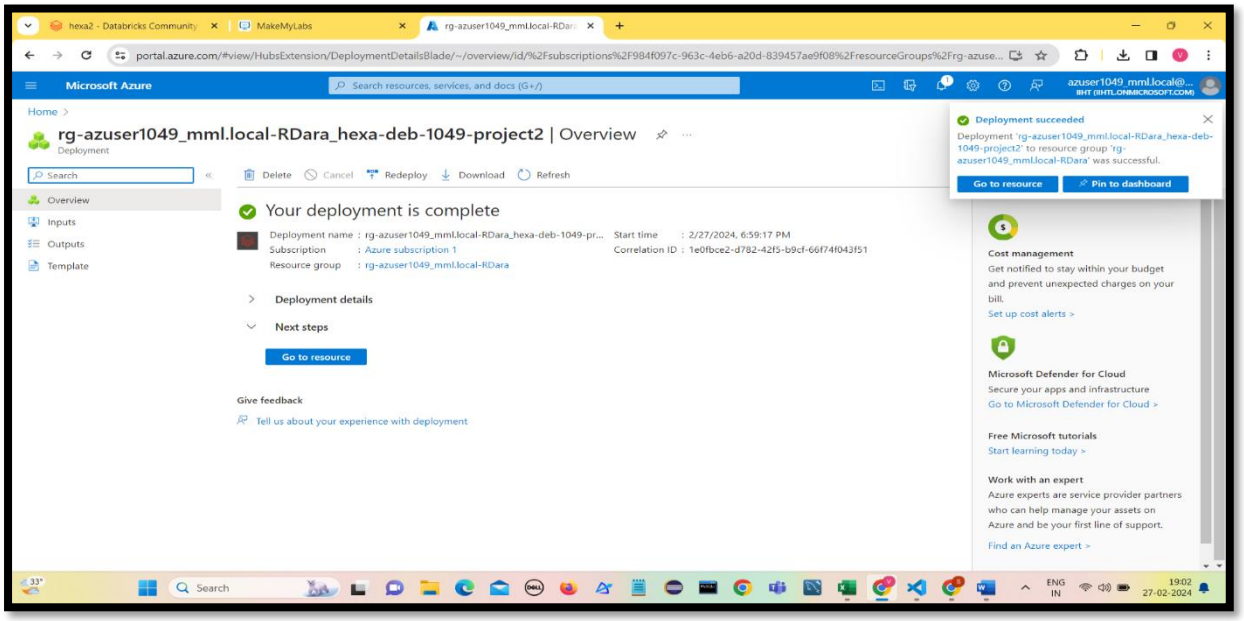


2) Create an Azure Databricks workspace by selecting Subscription, Resource Group, Workspace name(hexa-deb-1049-project2),Region,Pricing Tier, Click on **Review+Create.**

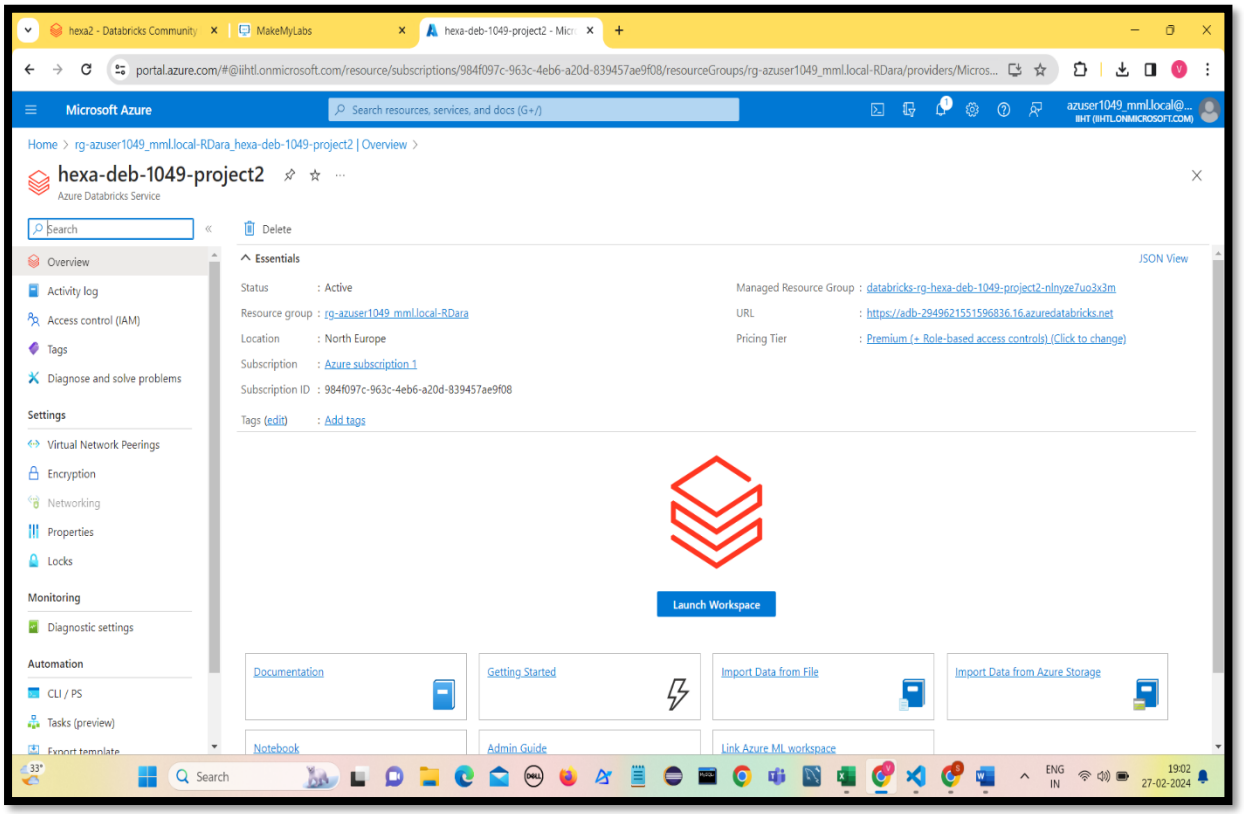




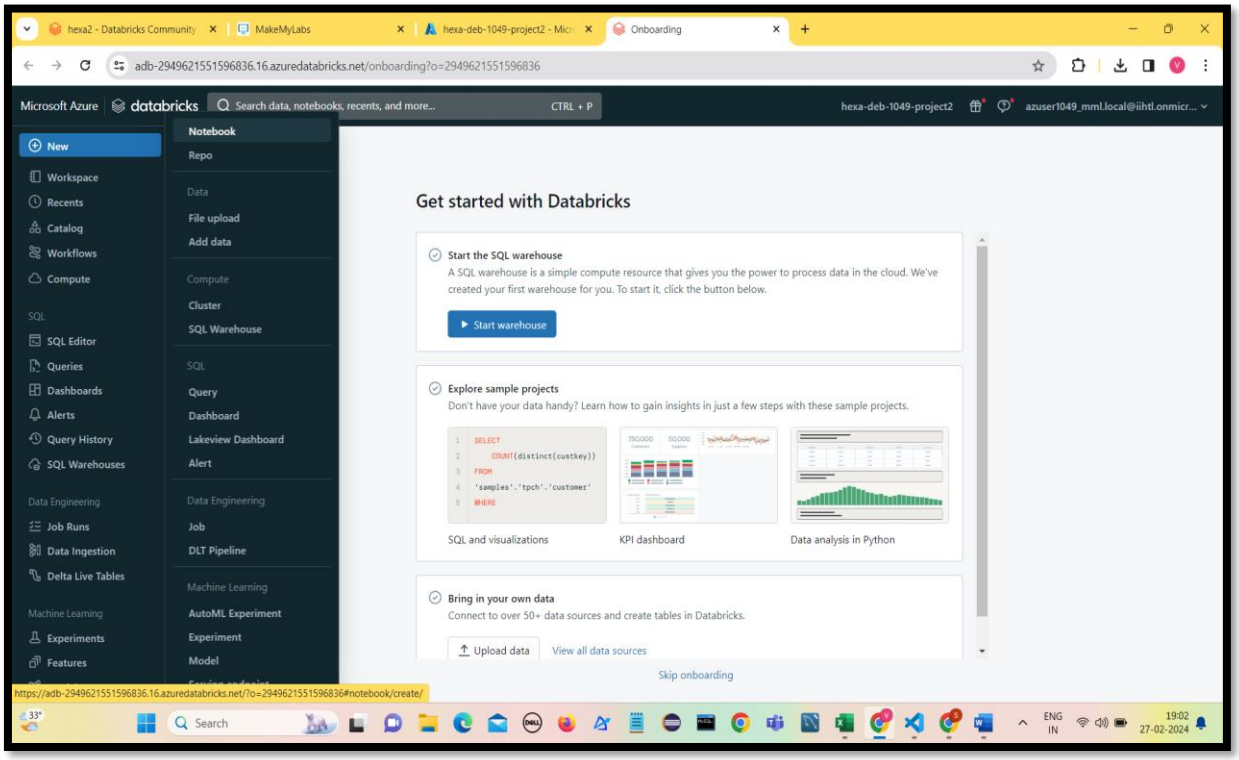
3) After the Validation is Succeeded , Click on Create.



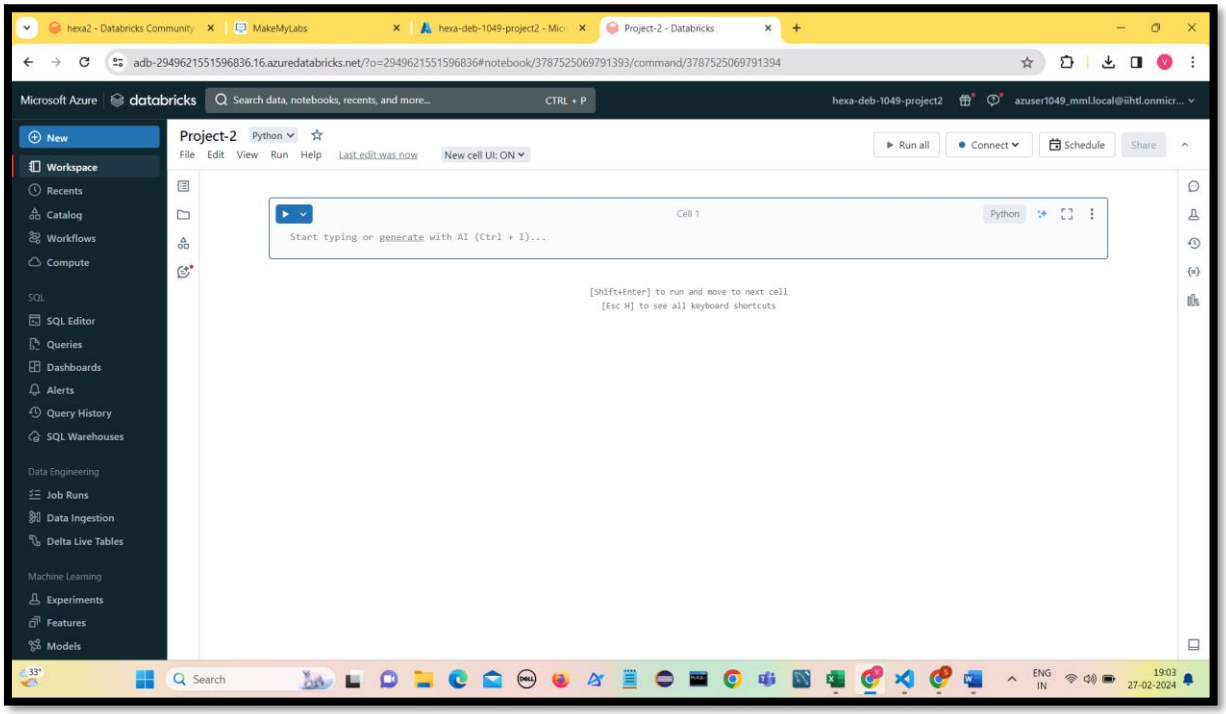
4) Once deployment is complete, Click on Go to resource.



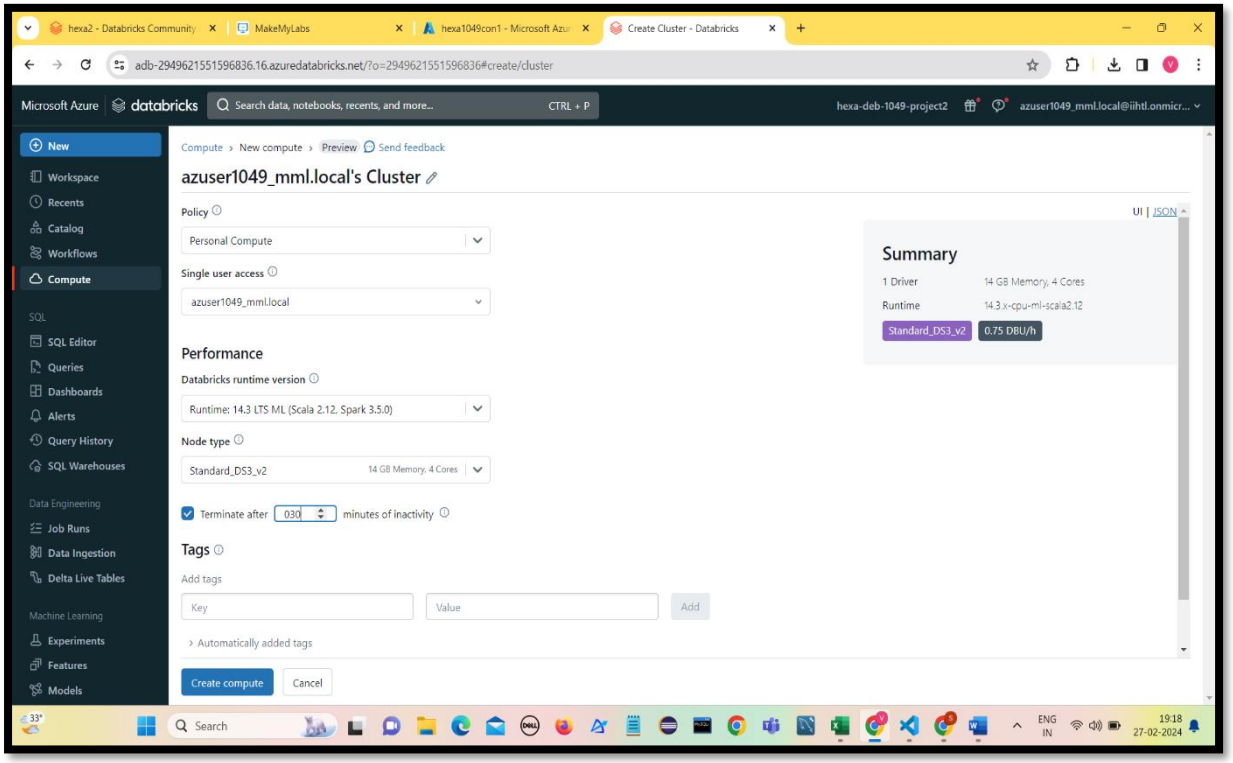
5) Click on Launch Workspace.

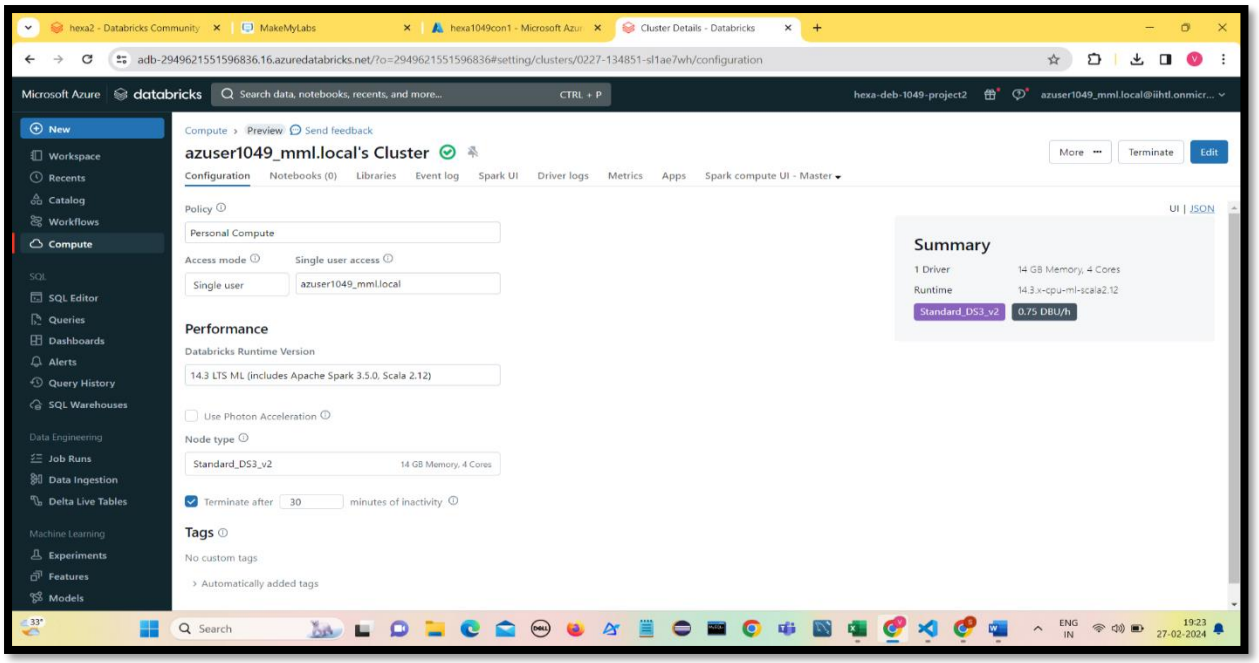


6) To create a Databricks notebook , Click New and click Notebook.



7) Create Compute by choosing Policy , Version, Node time etc., and Click on Create Compute,

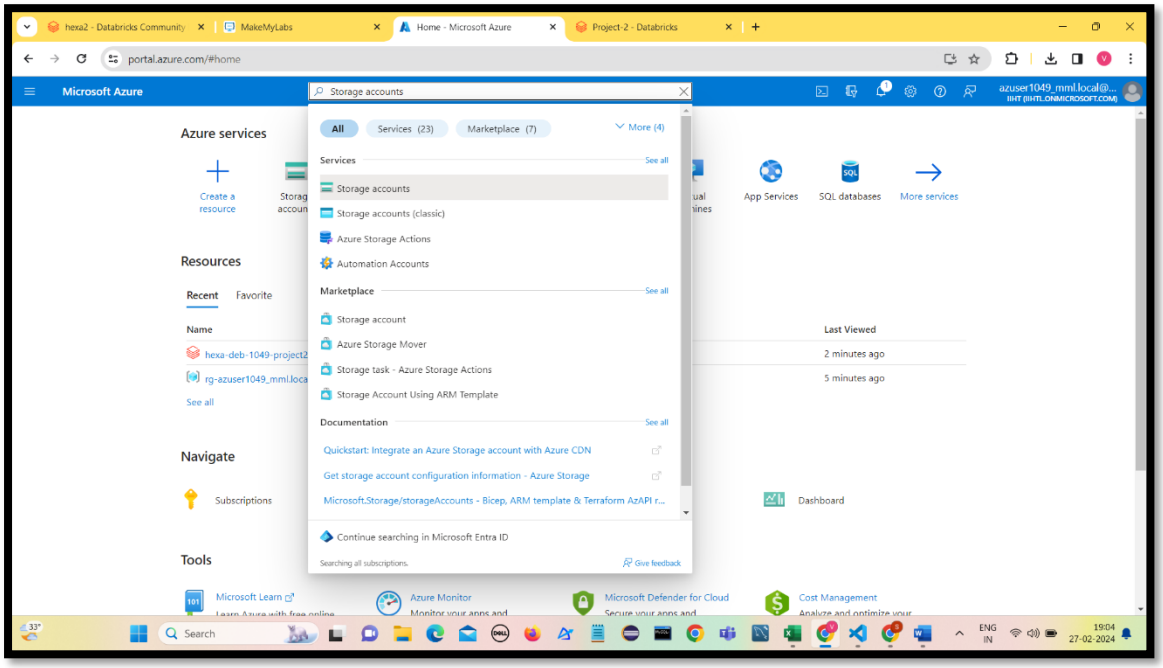




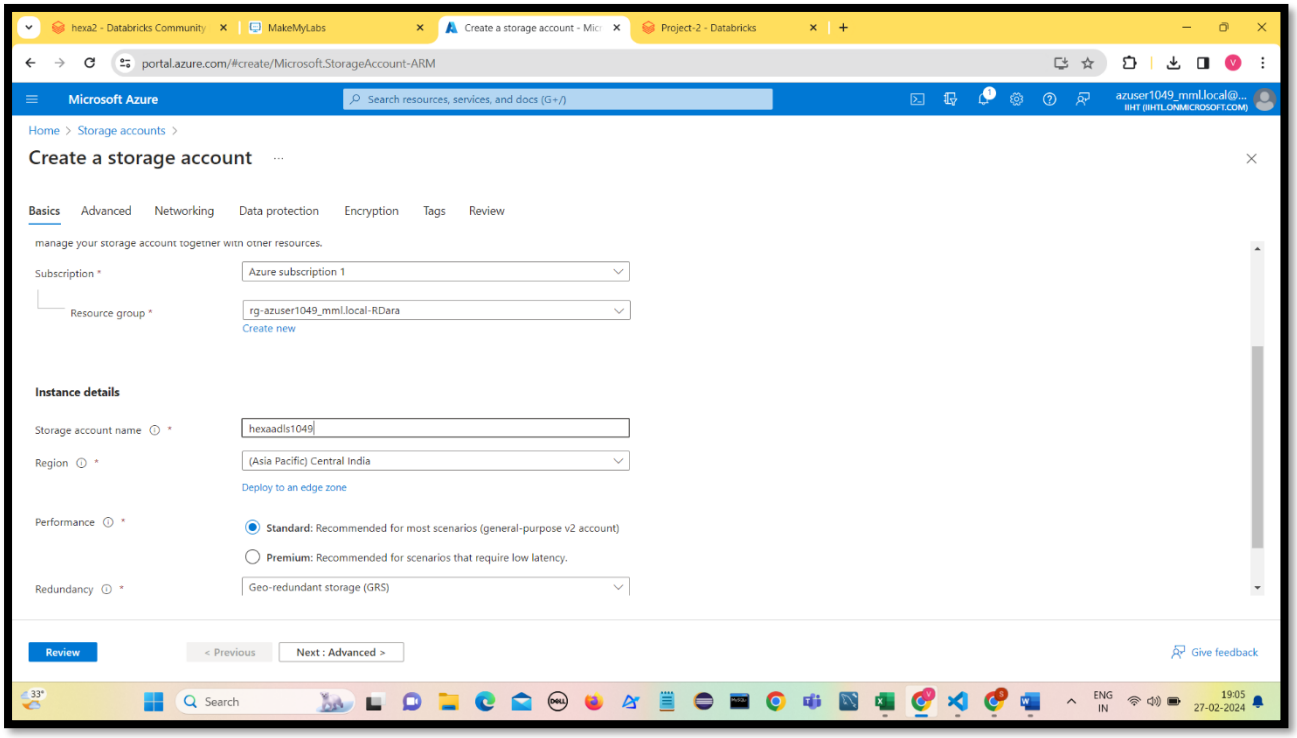
8) Once the Cluster is created, we can use it to run the PySpark Notebooks.

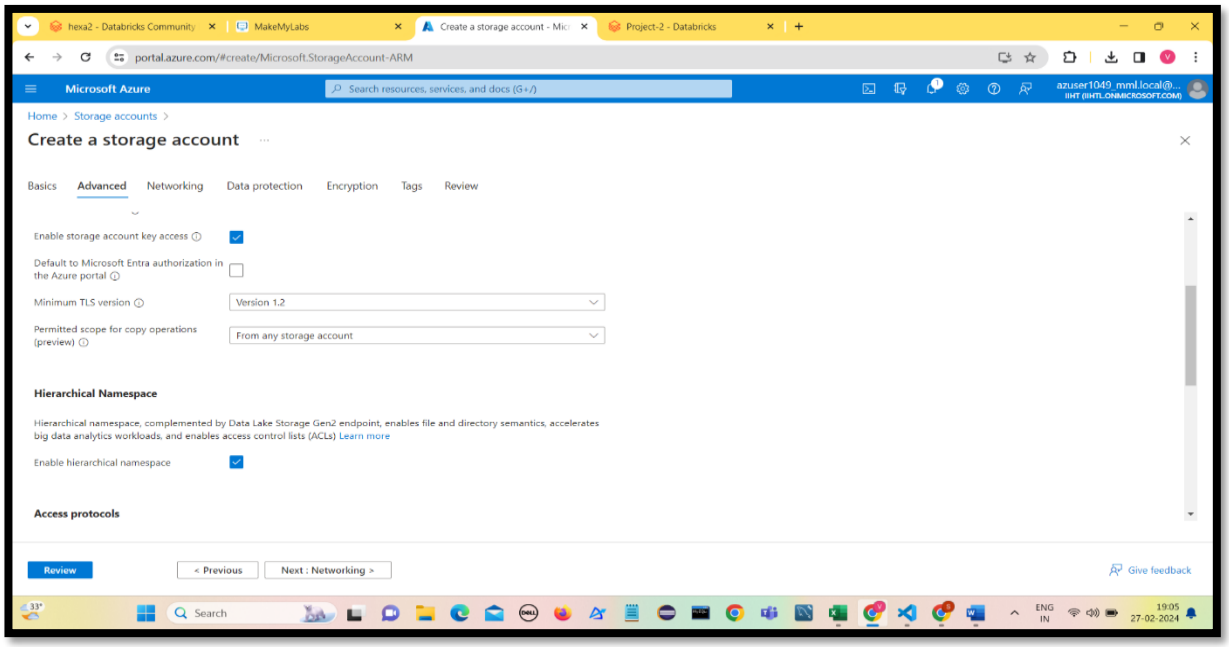
Creating ADLS Gen2 account (hexaadls1049)

9) Search for Storage accounts in Azure portal.

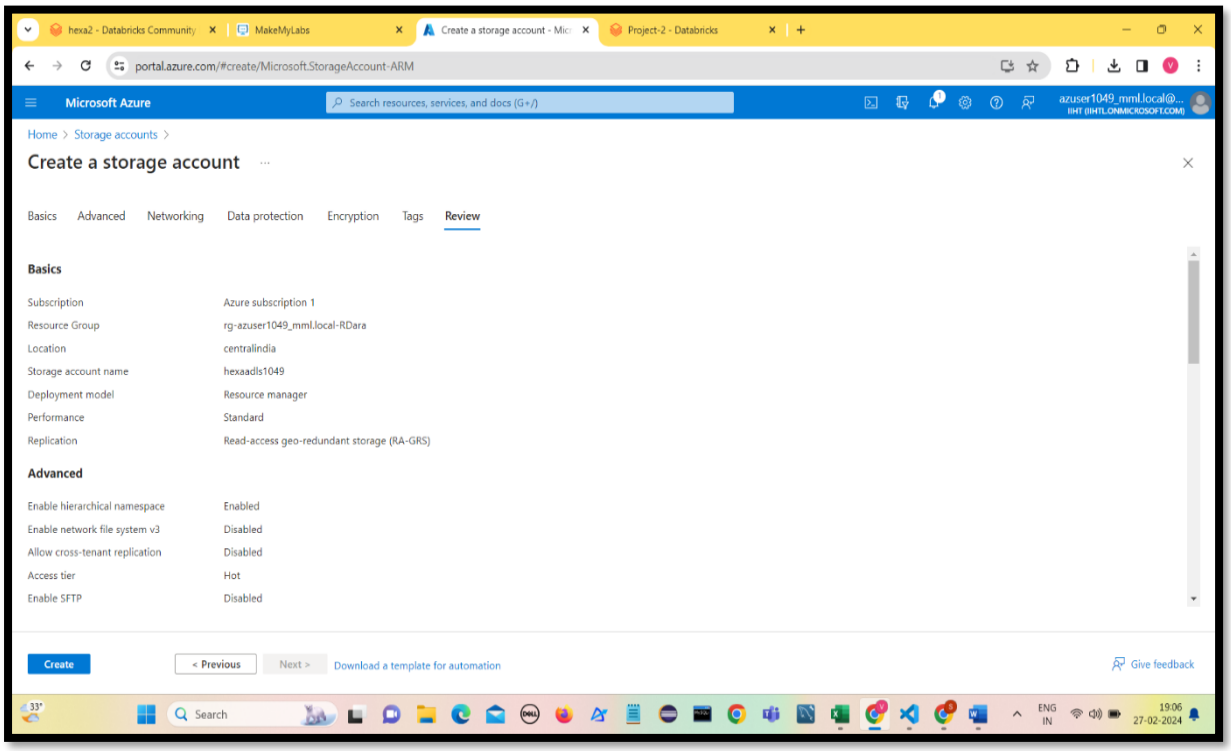


10) To create a ADLS account choose the Subscription, Storage account name (hexaadls1049), Region,Performance etc., and Click on Review.

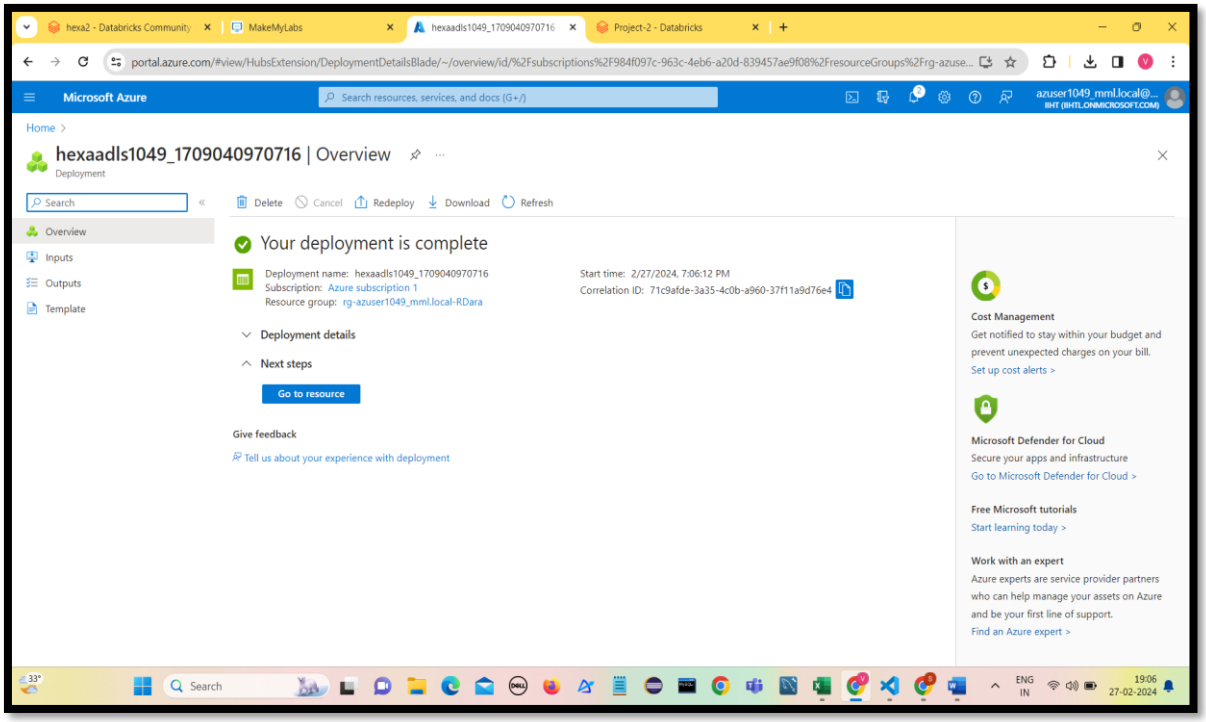




11) Enable Hierarchical Namespace in Advanced Settings and Click on Review.

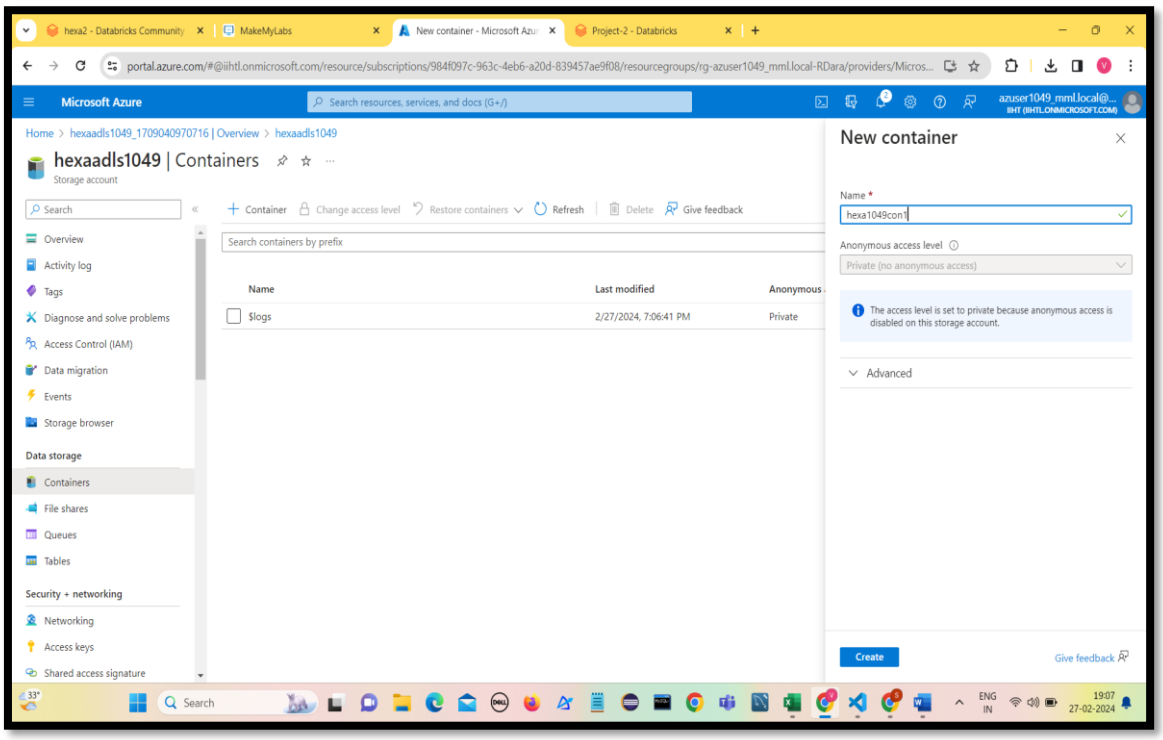


12) Click on Create and Deployment is in process.



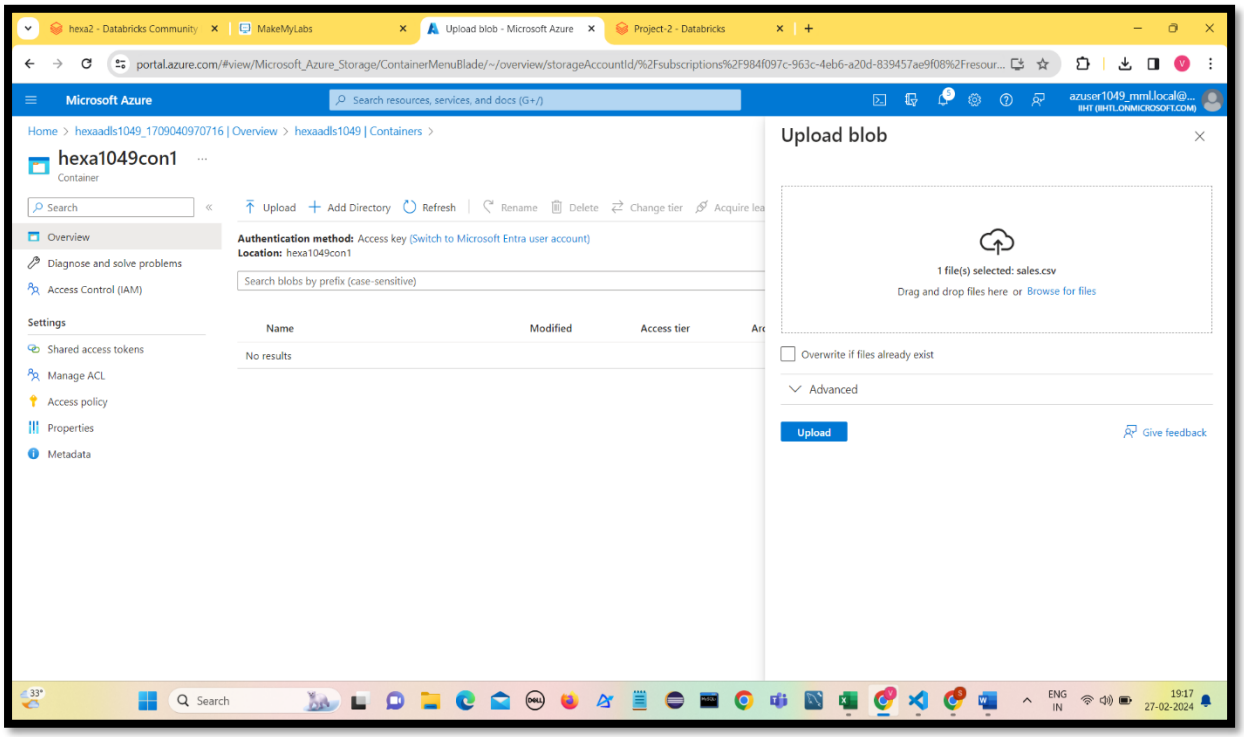
13) Click on resource.

14) Open the hexaadls1049 and Click on Containers and enter the Name of container.

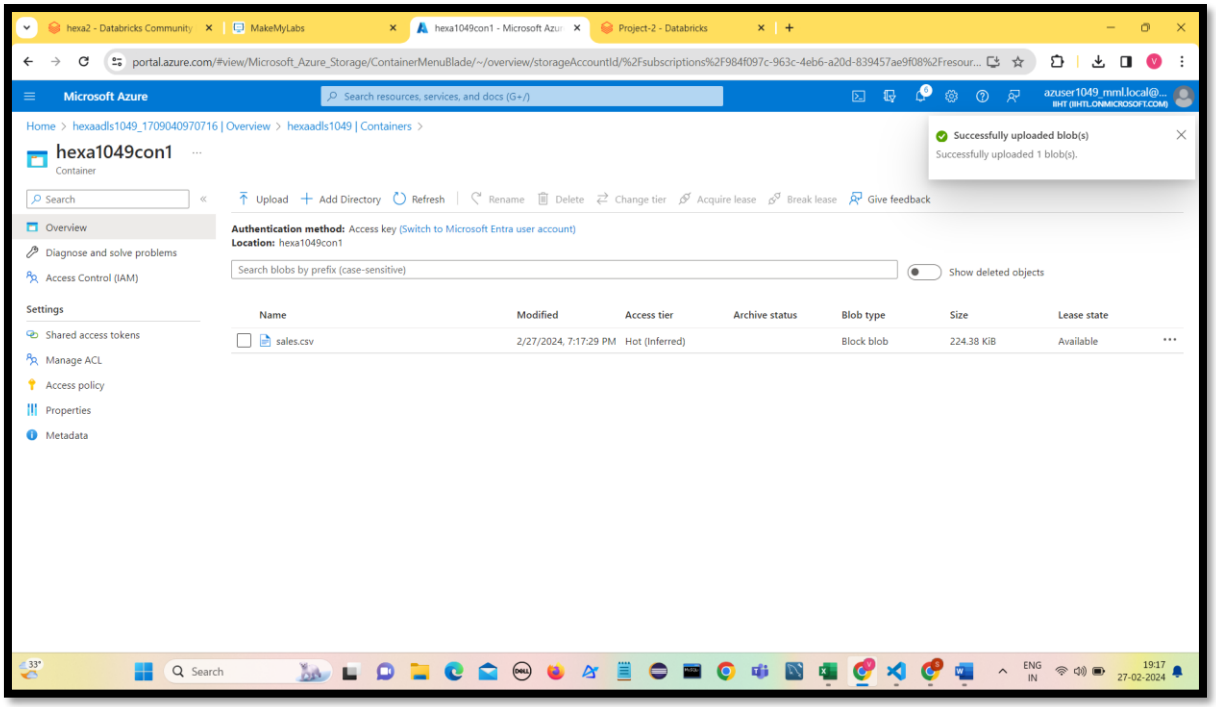


15) Click on Create.

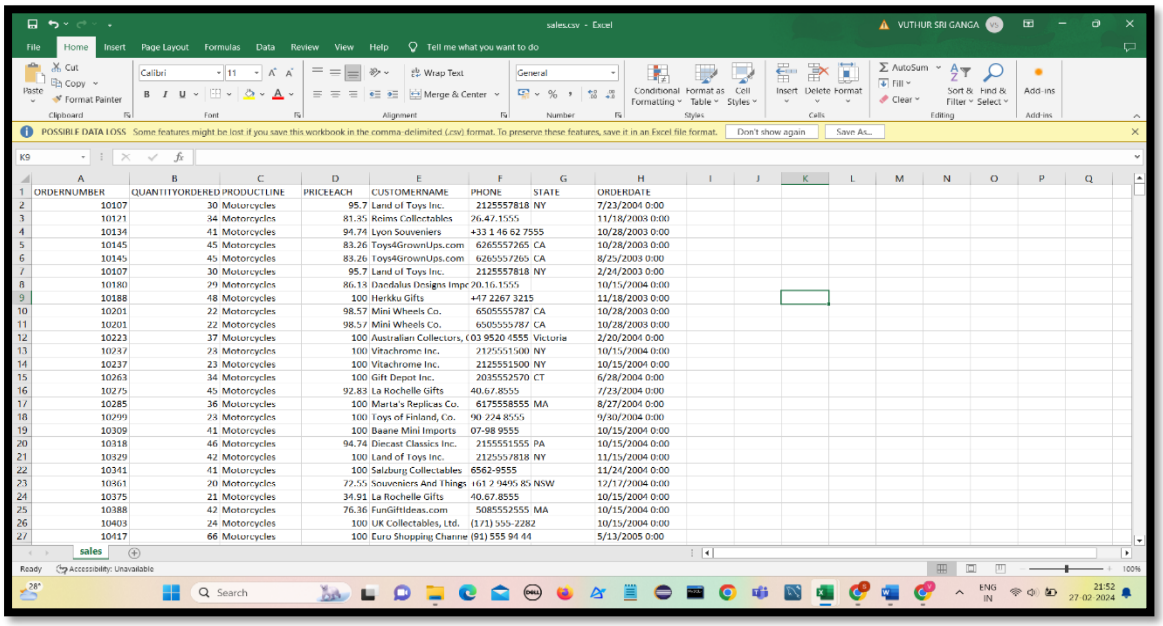
16) Select the file to upload into the container.



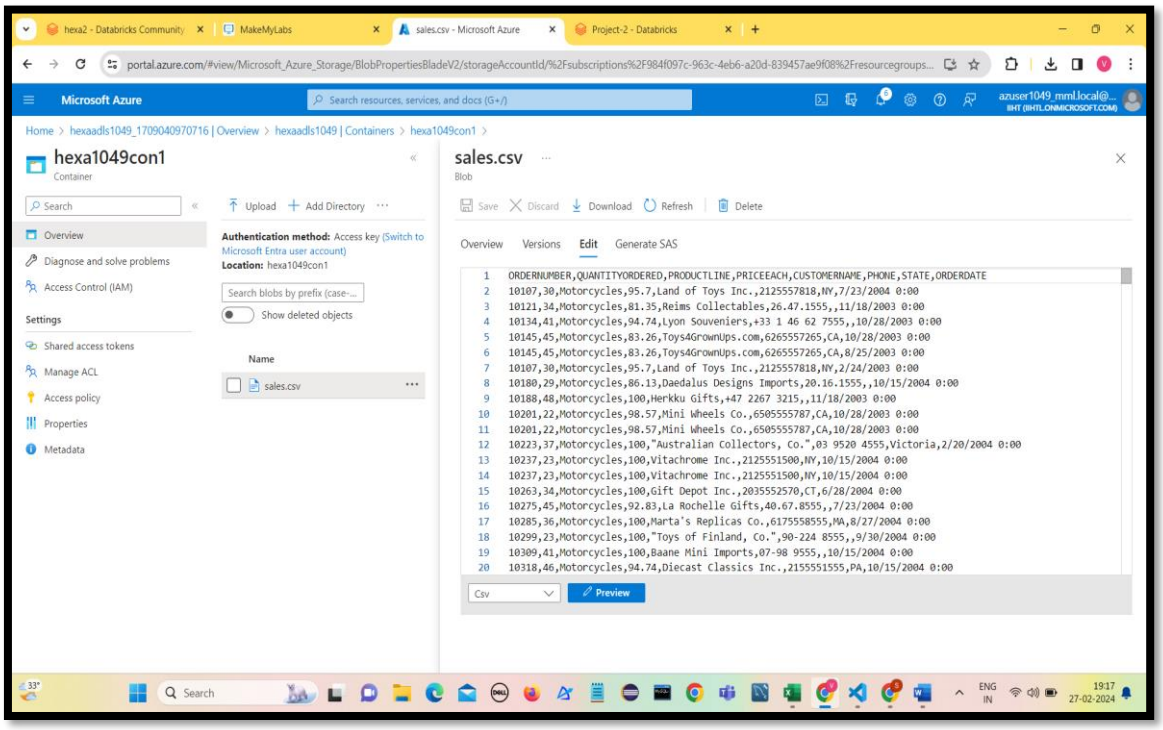
17) Click on Upload



18) sales.csv is added to the container.

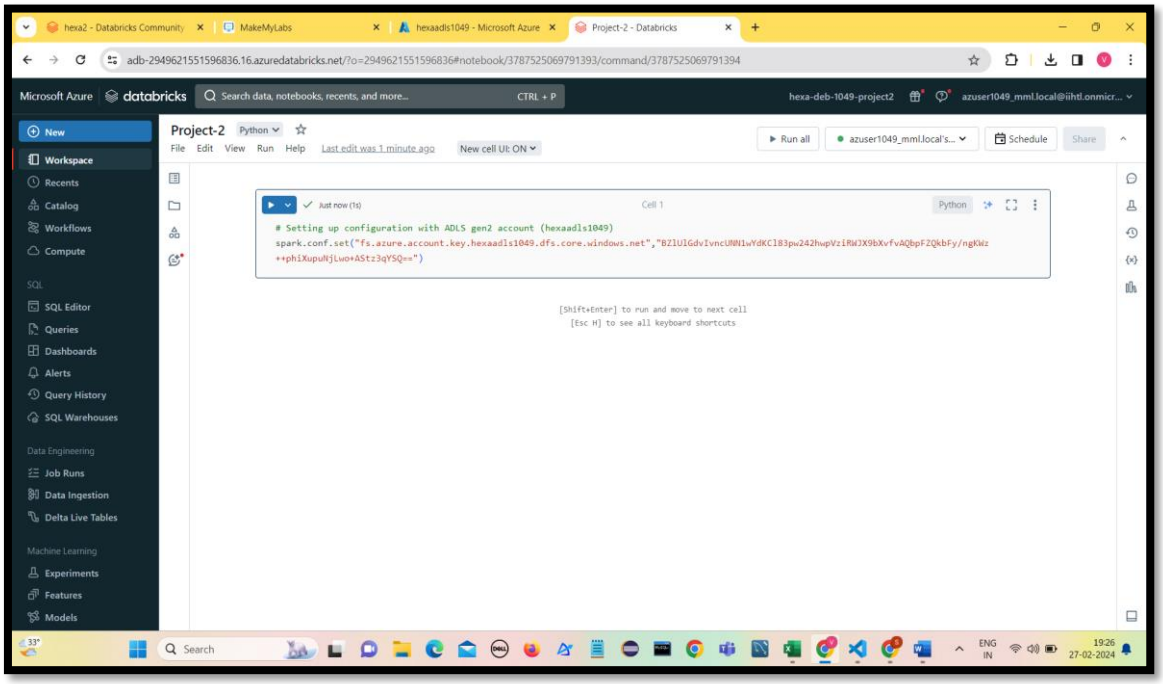


ORDERNUMBER	QUANTITYORDERED	PRODUCTLINE	PRICEEACH	CUSTOMERNAME	PHONE	STATE	ORDERDATE
10107	30	Motorcycles	95.7	Land of Toys Inc.	2125557818	NY	7/23/2004 0:00
10121	34	Motorcycles	81.35	Reims Collectables	26.47.1555		11/18/2003 0:00
10134	41	Motorcycles	94.74	Lyon Souveniers	+33 1 46 62 7555		10/28/2003 0:00
10145	45	Motorcycles	83.26	Toys4GrownUps.com	6265557265	CA	10/28/2003 0:00
10145	45	Motorcycles	83.26	Toys4GrownUps.com	6265557265	CA	8/25/2003 0:00
10107	30	Motorcycles	95.7	Land of Toys Inc.	2125557818	NY	2/24/2003 0:00
10180	29	Motorcycles	86.13	Daedalus Designs Impc	20.16.1555		10/15/2004 0:00
10188	48	Motorcycles	100	Herkku Gifts	+47 2267 3215		11/18/2003 0:00
10201	22	Motorcycles	98.57	Mini Wheels Co.	6505555787	CA	10/28/2003 0:00
10201	22	Motorcycles	98.57	Mini Wheels Co.	6505555787	CA	10/28/2003 0:00
10223	37	Motorcycles	100	Australian Collectors, C	03 9520 4555	Victoria	2/20/2004 0:00
10237	23	Motorcycles	100	Vitachrome Inc.	2125551500	NY	10/15/2004 0:00
10237	23	Motorcycles	100	Vitachrome Inc.	2125551500	NY	10/15/2004 0:00
10263	34	Motorcycles	100	Gift Depot Inc.	2035552570	CT	6/28/2004 0:00
10275	45	Motorcycles	92.83	La Rochelle Gifts	40.67.8555		7/23/2004 0:00
10285	36	Motorcycles	100	Marta's Replicas Co.	6175558555	MA	8/27/2004 0:00
10299	23	Motorcycles	100	Toys of Finland, Co.	90 224 8555		9/30/2004 0:00
10309	41	Motorcycles	100	Baane Mini Imports	07-98 9555		10/15/2004 0:00
10318	46	Motorcycles	94.74	Diecast Classics Inc.	2155551555	PA	10/15/2004 0:00
10329	42	Motorcycles	100	Land of Toys Inc.	2125557818	NY	11/15/2004 0:00
10341	41	Motorcycles	100	Salzburg Collectables	6562-9555		11/24/2004 0:00
10361	20	Motorcycles	72.55	Souveniers And Things	161 2 9495 85	NSW	12/17/2004 0:00
10375	21	Motorcycles	34.91	La Rochelle Gifts	40.67.8555		10/15/2004 0:00
10388	42	Motorcycles	76.36	FunGiftIdeas.com	5085552555	MA	10/15/2004 0:00
10403	24	Motorcycles	100	UK Collectables, Ltd.	(171) 555-2282		10/15/2004 0:00
10417	66	Motorcycles	100	Euro Shopping Channe	(91) 555 94 44		5/13/2005 0:00

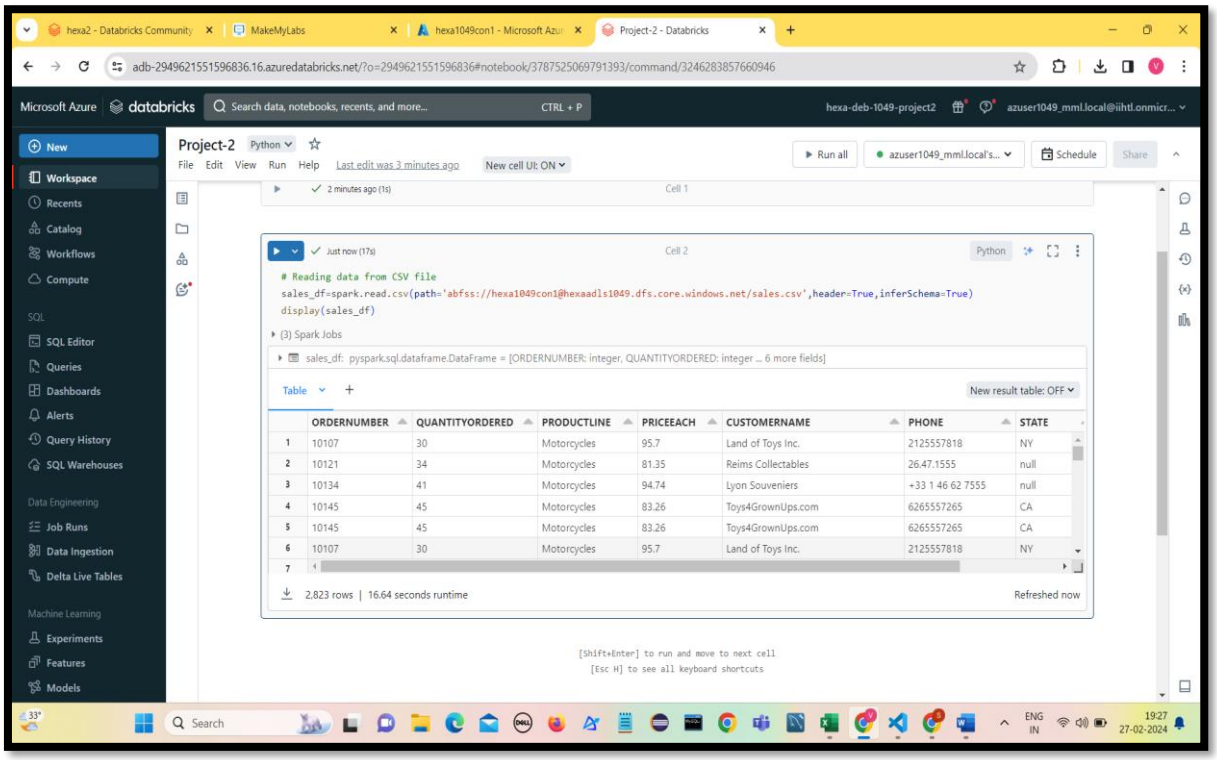


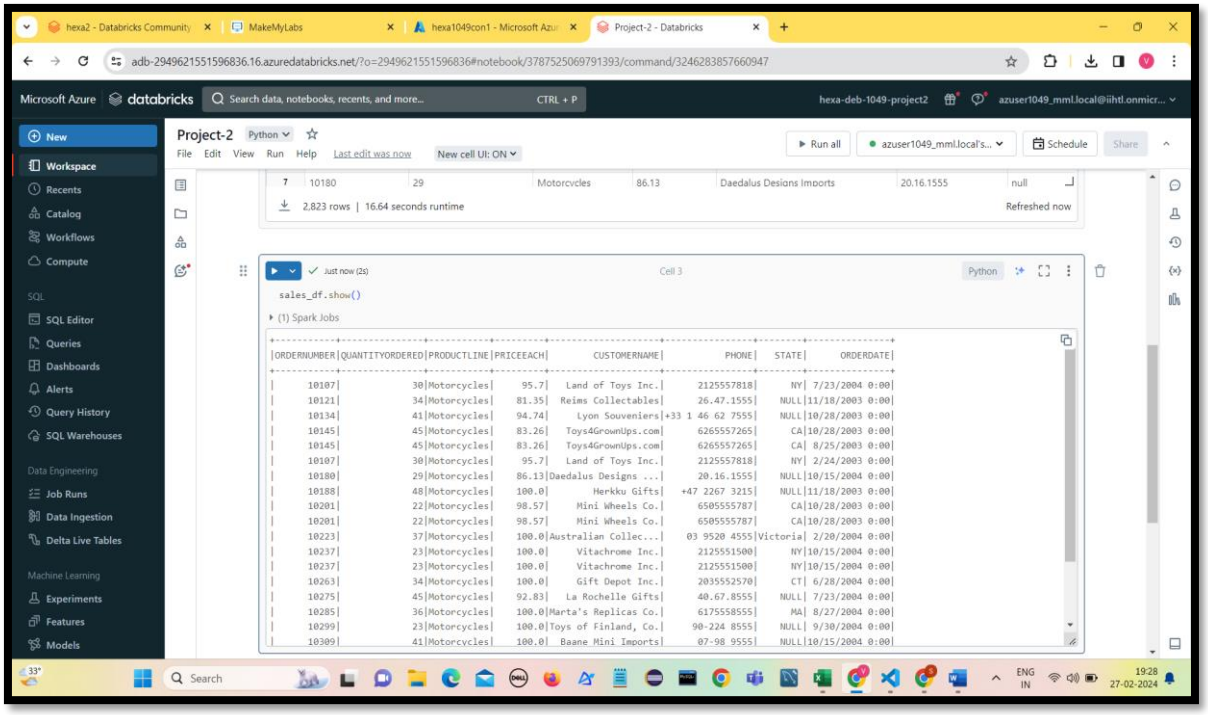
```
ORDERNUMBER,QUANTITYORDERED,PRODUCTLINE,PRICEEACH,CUSTOMERNAME,PHONE,STATE,ORDERDATE
10107,30,Motorcycles,95.7,Land of Toys Inc.,2125557818,NY,7/23/2004 0:00
10121,34,Motorcycles,81.35,Reims Collectables,26.47.1555,,11/18/2003 0:00
10134,41,Motorcycles,94.74,Lyon Souveniers,+33 1 46 62 7555,,10/28/2003 0:00
10145,45,Motorcycles,83.26,Toys4GrownUps.com,6265557265,CA,10/28/2003 0:00
10145,45,Motorcycles,83.26,Toys4GrownUps.com,6265557265,CA,8/25/2003 0:00
10107,30,Motorcycles,95.7,Land of Toys Inc.,2125557818,NY,2/24/2003 0:00
10180,29,Motorcycles,86.13,Daedalus Designs Imports,20.16.1555,,10/15/2004 0:00
10188,48,Motorcycles,100,Herkku Gifts,+47 2267 3215,,11/18/2003 0:00
10201,22,Motorcycles,98.57,Mini Wheels Co.,6505555787,CA,10/28/2003 0:00
10201,22,Motorcycles,98.57,Mini Wheels Co.,6505555787,CA,10/28/2003 0:00
10223,37,Motorcycles,100,"Australian Collectors, Co.",03 9520 4555,Victoria,2/20/2004 0:00
10237,23,Motorcycles,100,Vitachrome Inc.,2125551500,NY,10/15/2004 0:00
10237,23,Motorcycles,100,Vitachrome Inc.,2125551500,NY,10/15/2004 0:00
10263,34,Motorcycles,100,Gift Depot Inc.,2035552570,CT,6/28/2004 0:00
10275,45,Motorcycles,92.83,La Rochelle Gifts,40.67.8555,,7/23/2004 0:00
10285,36,Motorcycles,100,Marta's Replicas Co.,6175558555,MA,8/27/2004 0:00
10299,23,Motorcycles,100,"Toys of Finland, Co.",90-224 8555,,9/30/2004 0:00
10309,41,Motorcycles,100,Baane Mini Imports,07-98 9555,,10/15/2004 0:00
10318,46,Motorcycles,94.74,Diecast Classics Inc.,2155551555,PA,10/15/2004 0:00
```

19) Setting up configuration with ADLS gen2 account (hexaadls1049)



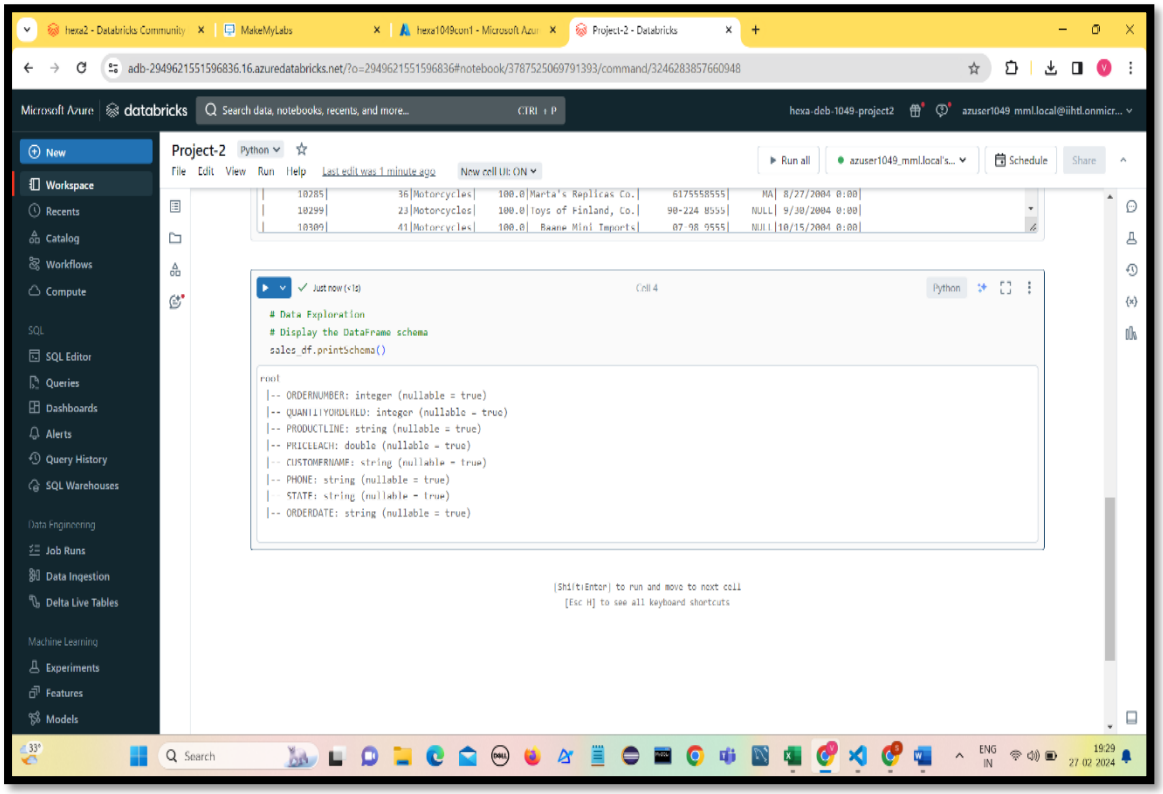
20) Reading data from sales.csv file using read.csv



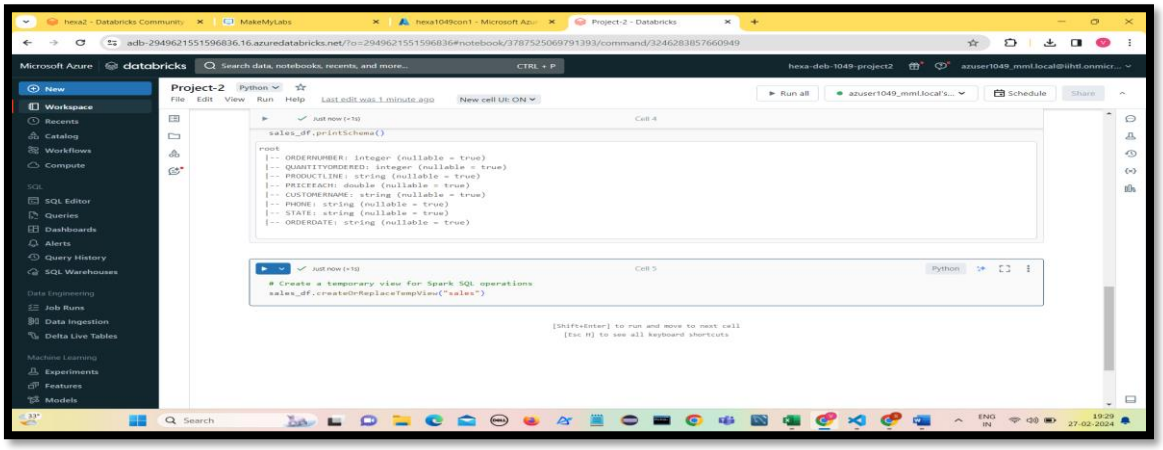


Data Exploration

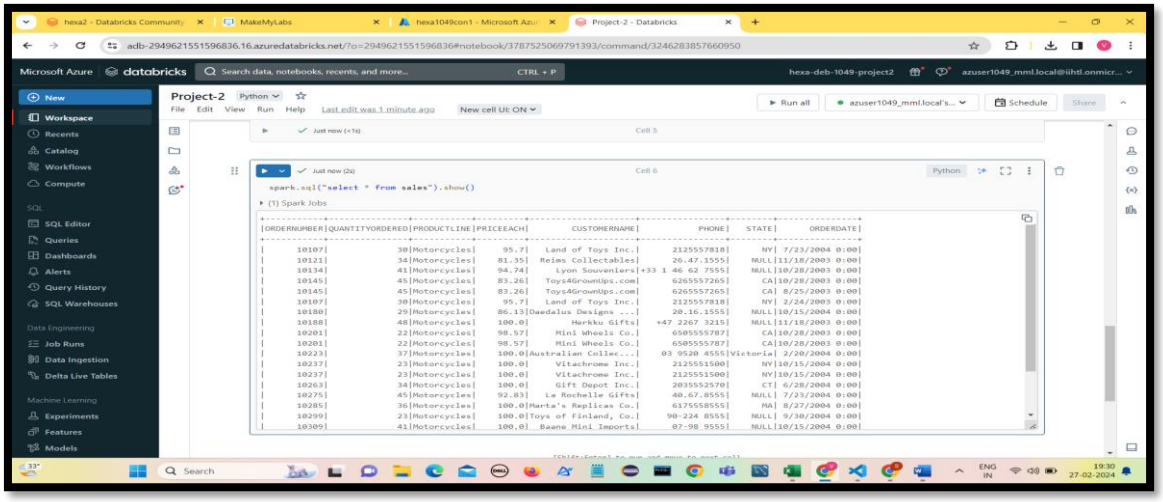
Display the DataFrame Schema.



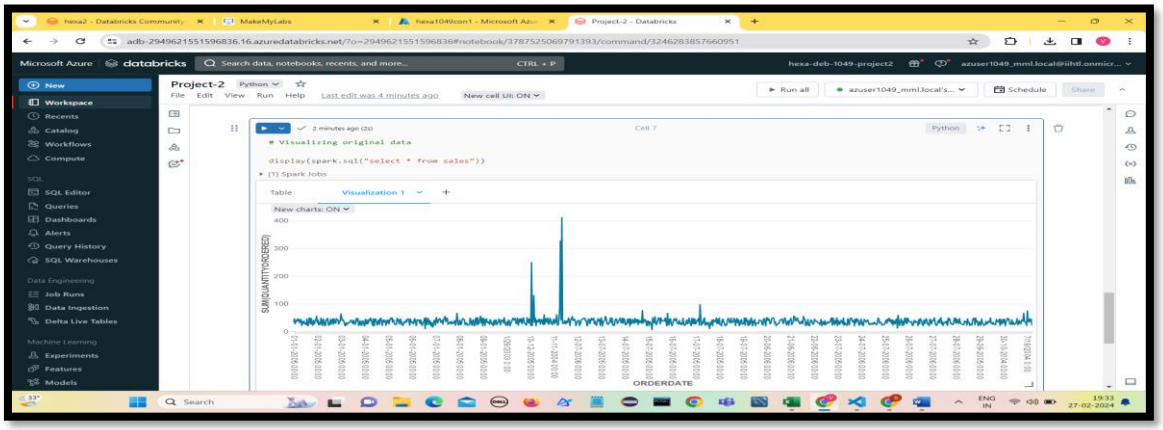
Create a temporary view for Spark SQL Operations.



Display the details of sales view.

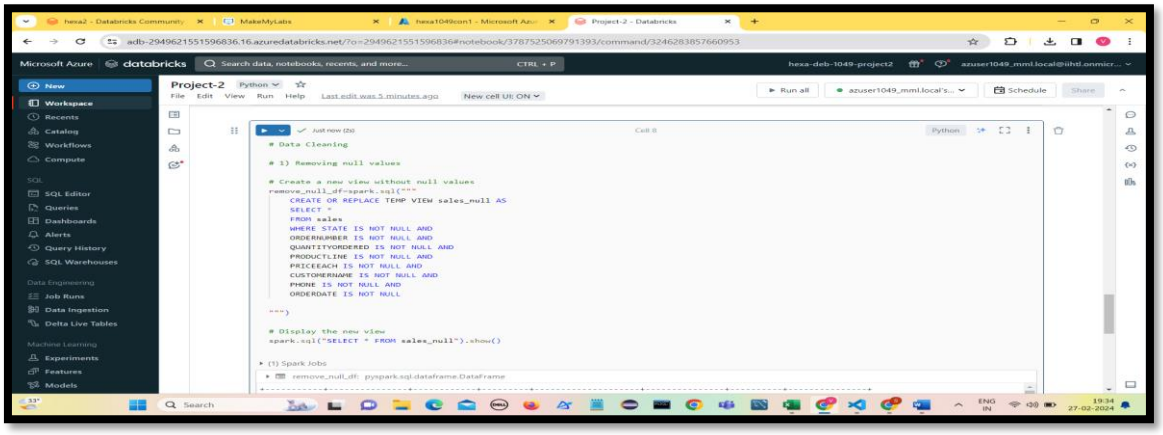


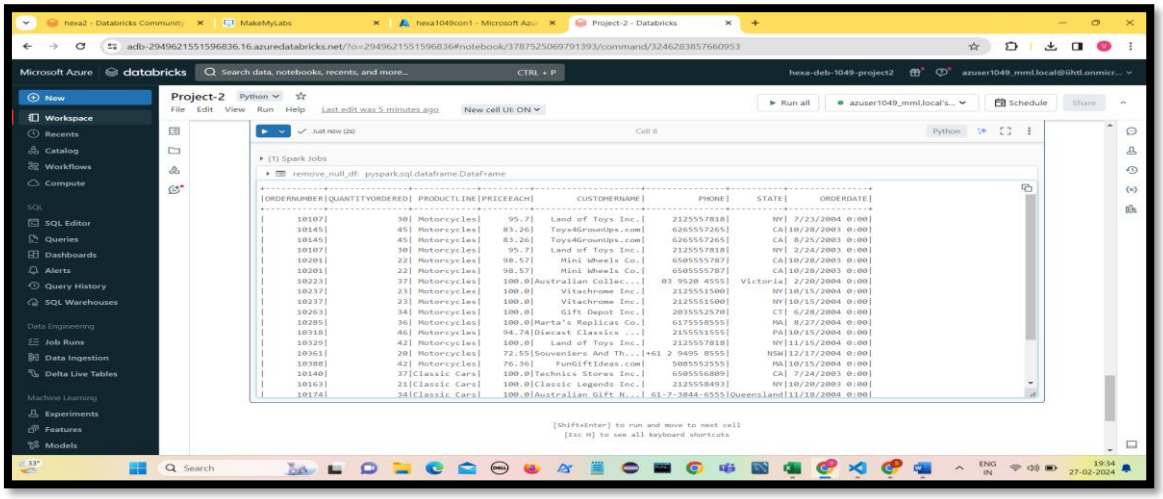
Visualizing original data using Line Chart.



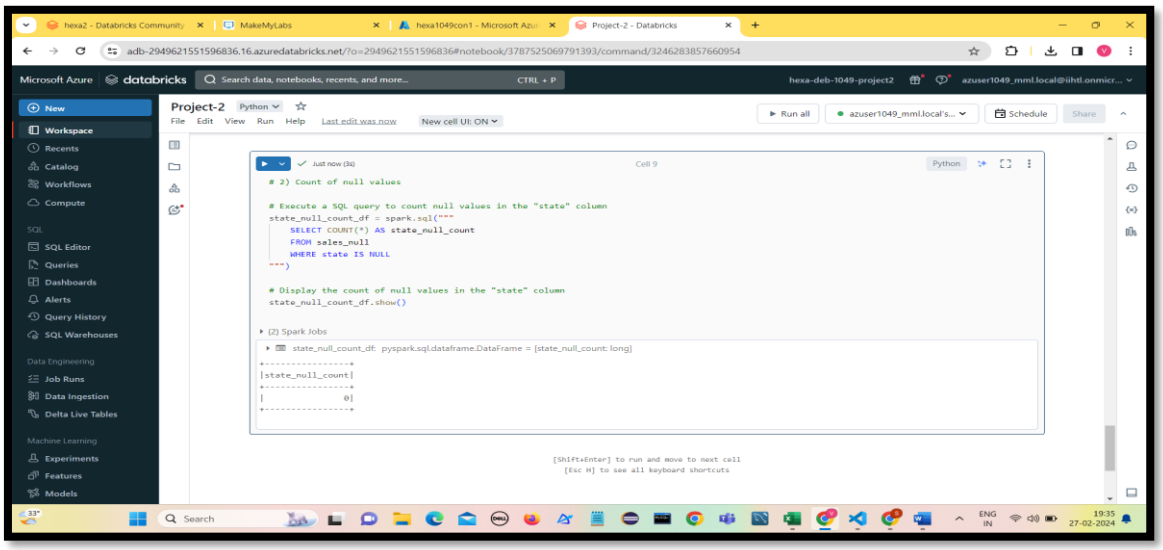
Data Cleaning

Removing null values

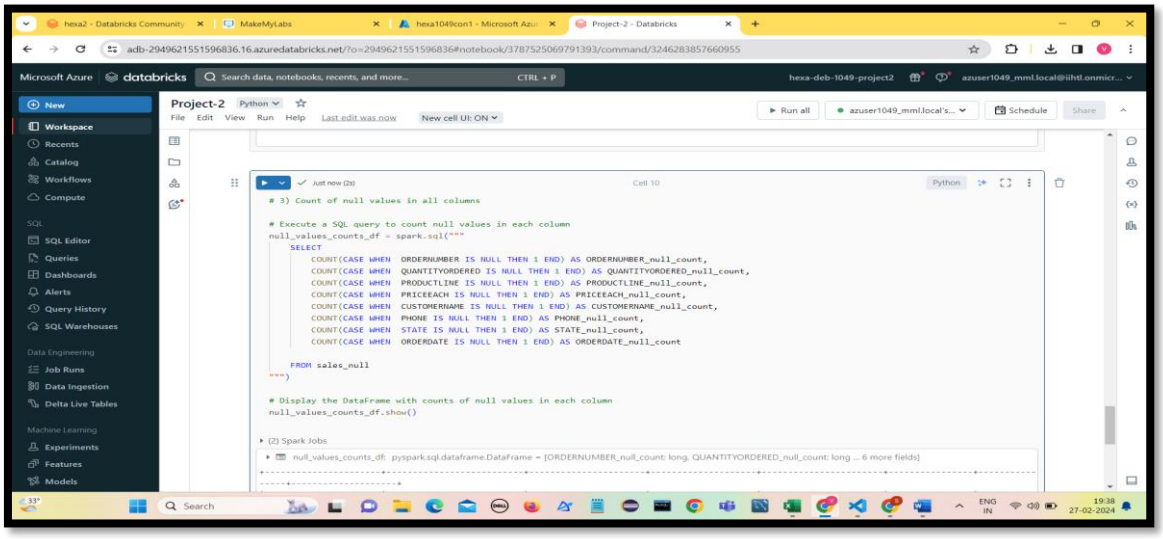


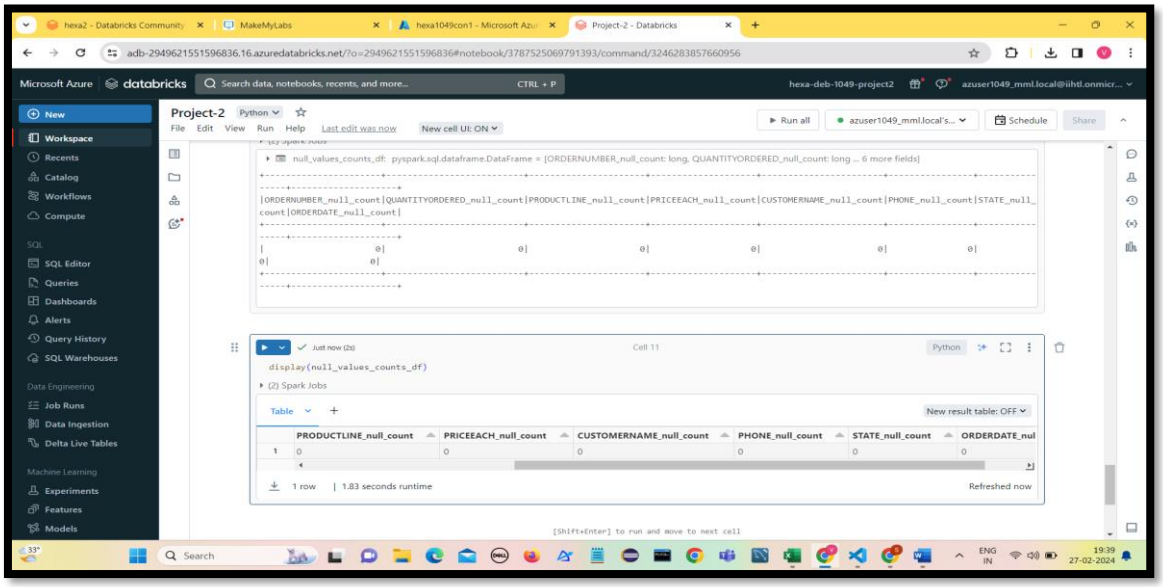


Count of null values in State column.

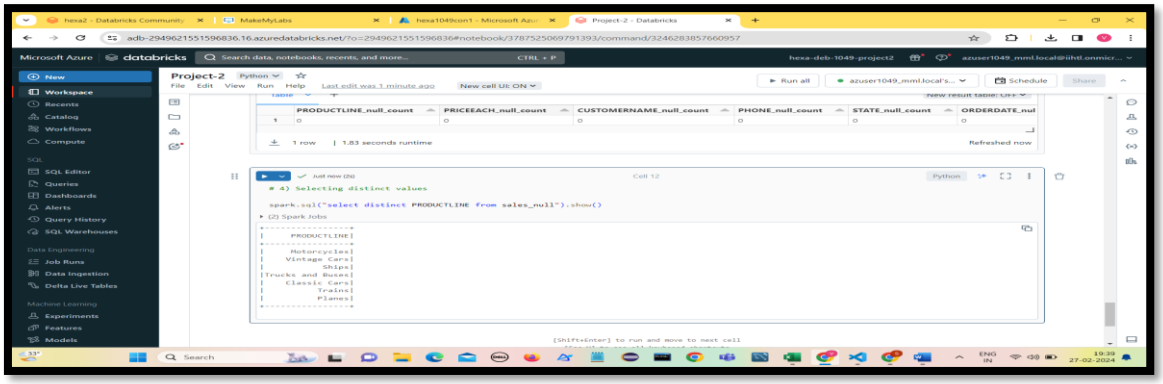


Count of Null values in all columns.

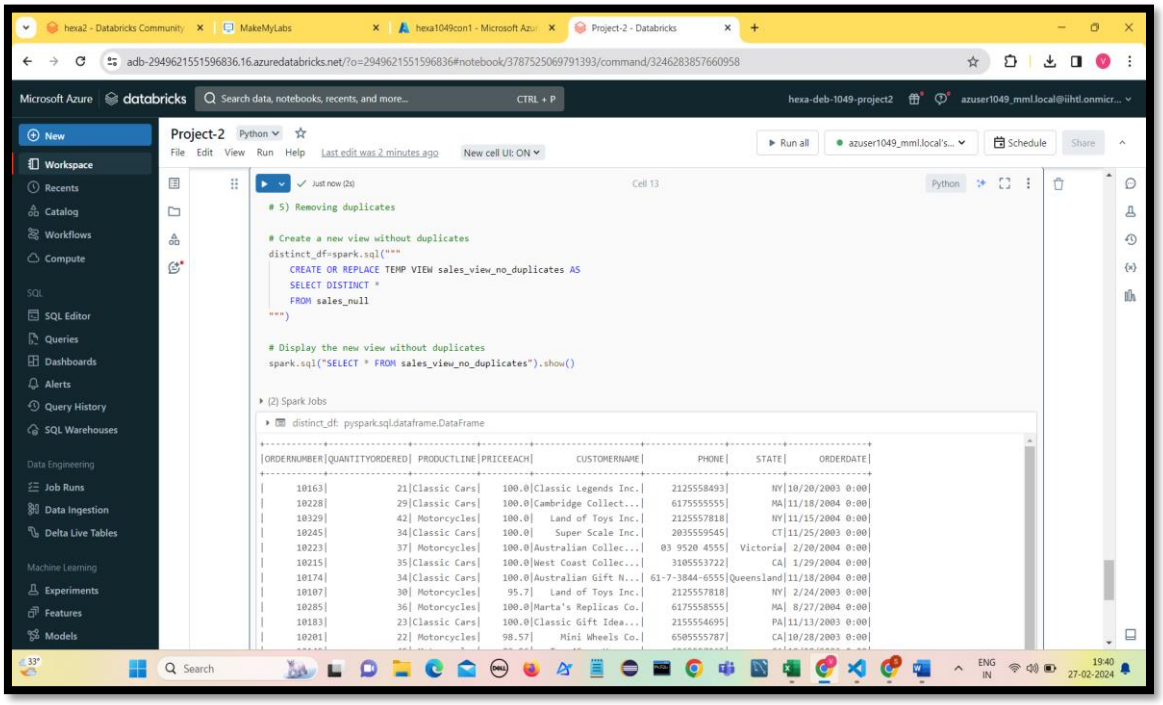




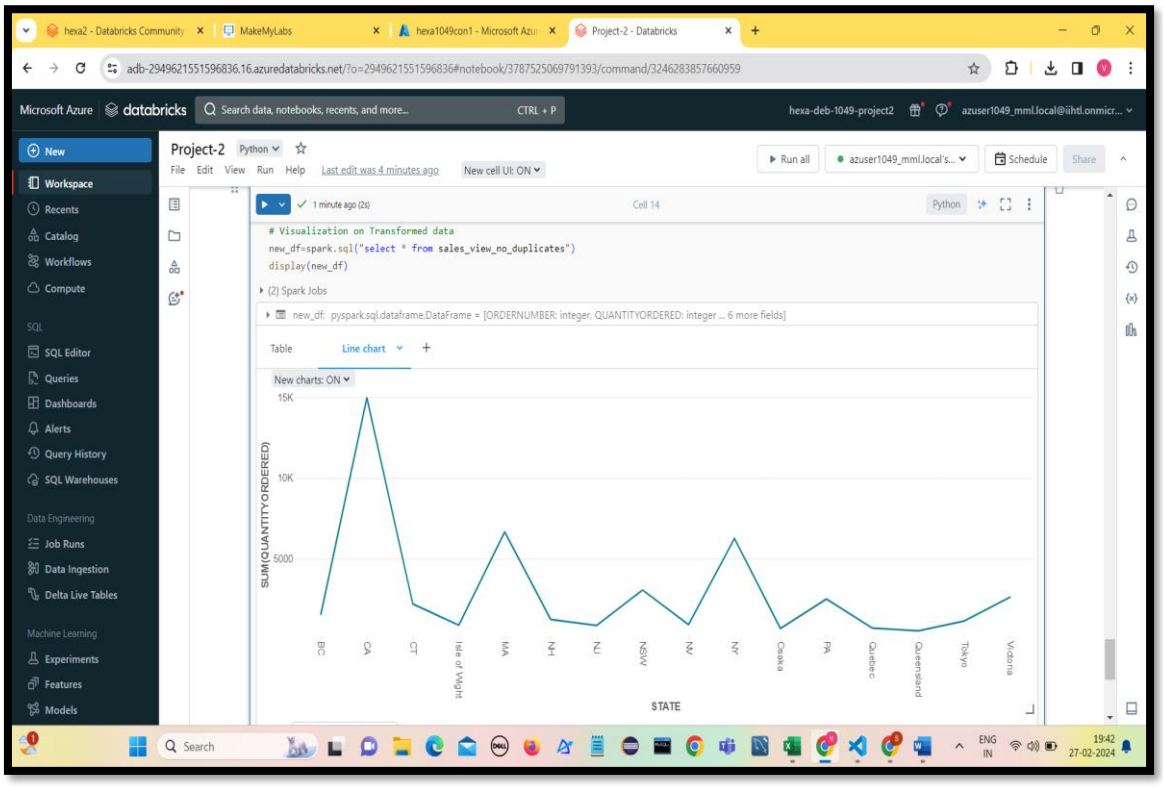
Selecting distinct PRODUCTLINE values



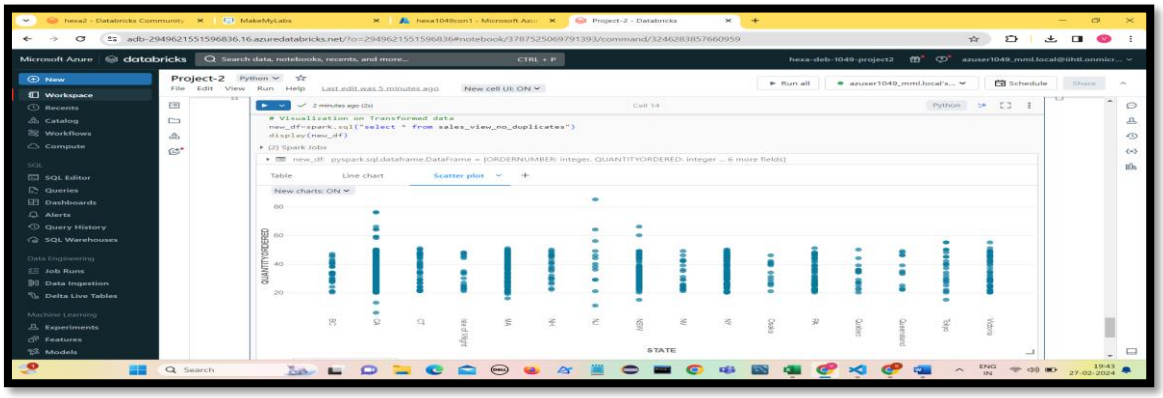
Removing duplicates from all columns



Visualization on Transformed data (Line chart)

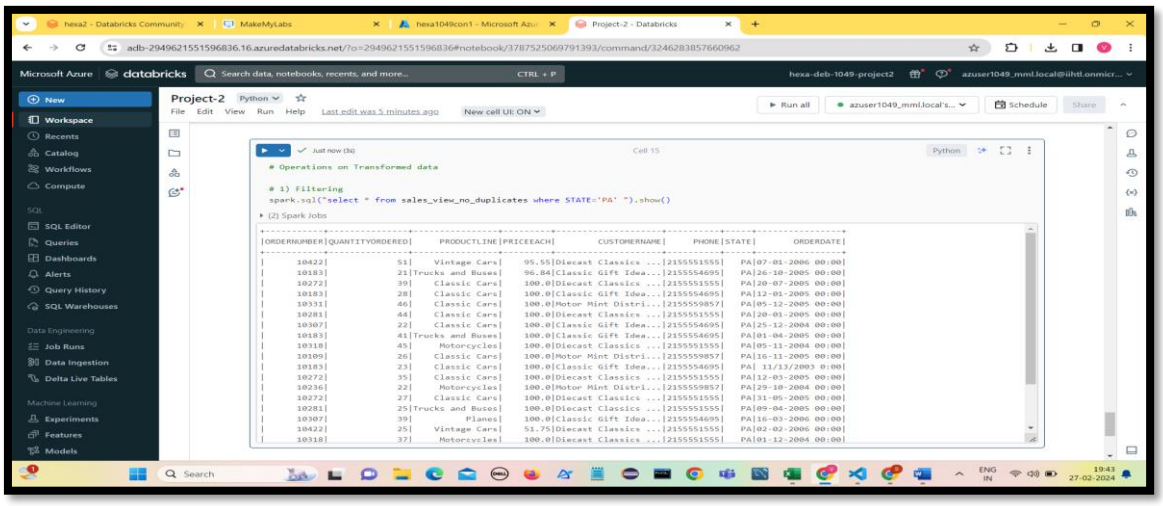


Scatter plot

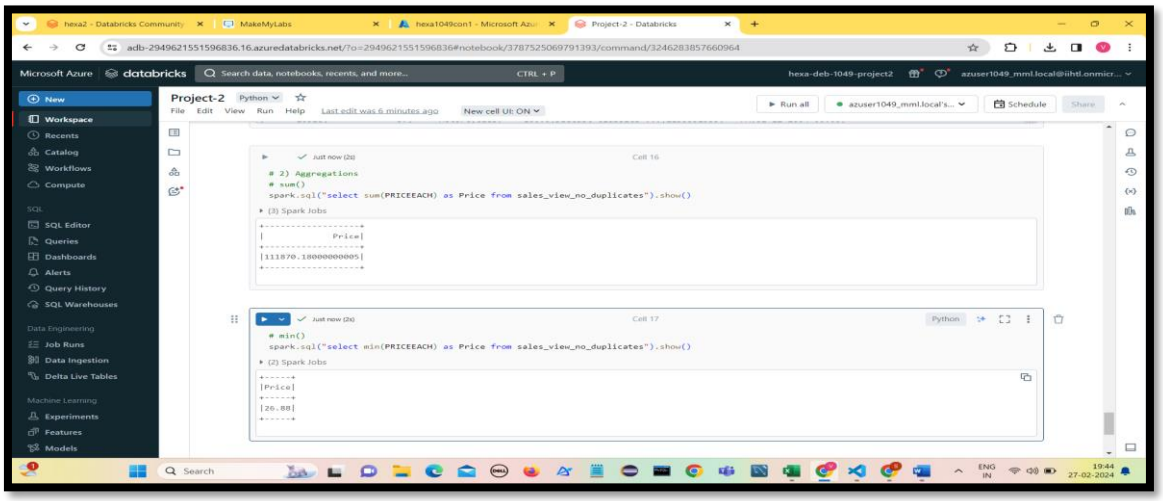


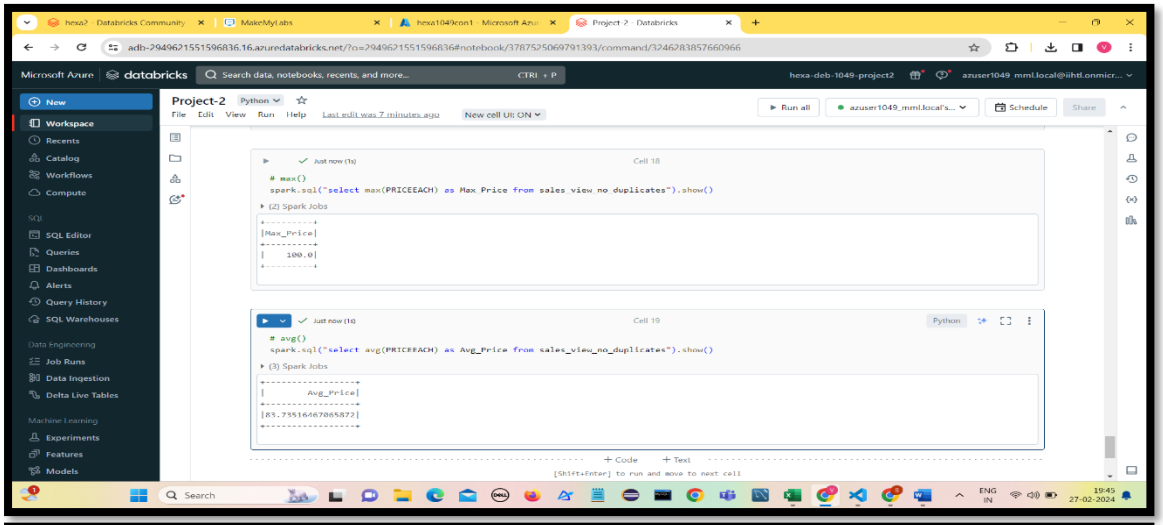
Operations on Transformed data

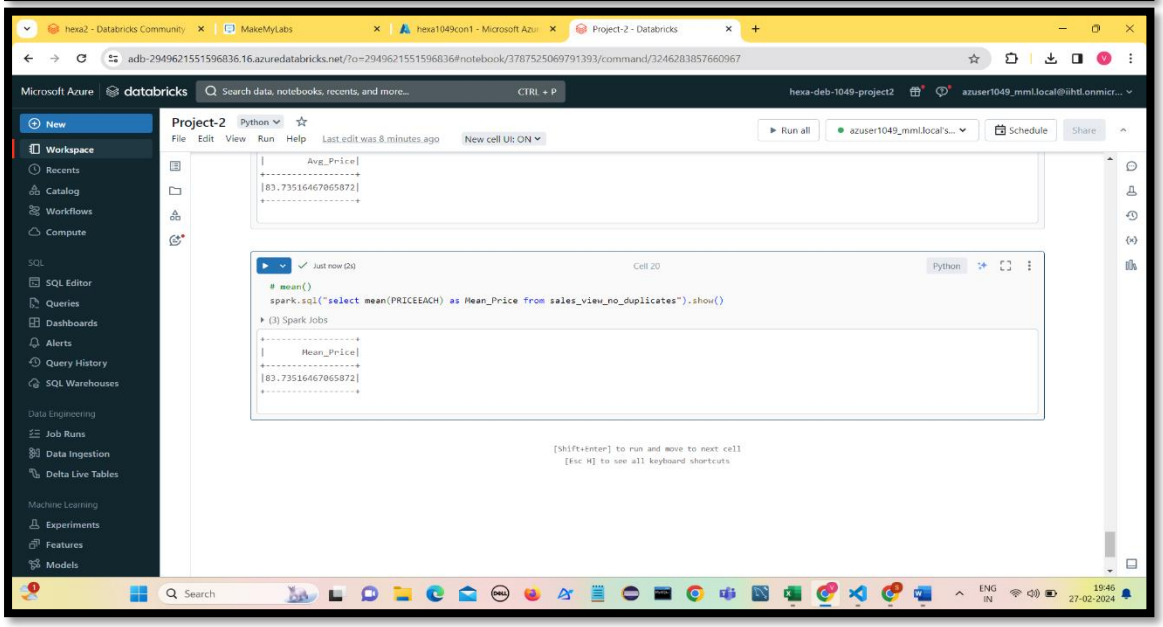
1) Filtering



2) Aggregations







1) Anomalies Detection (Calculating z-scores for each value)

The Z-score is a statistical measure that indicates how many standard deviations a data point is from the mean.

$Z = (X-\mu)/\sigma$

where

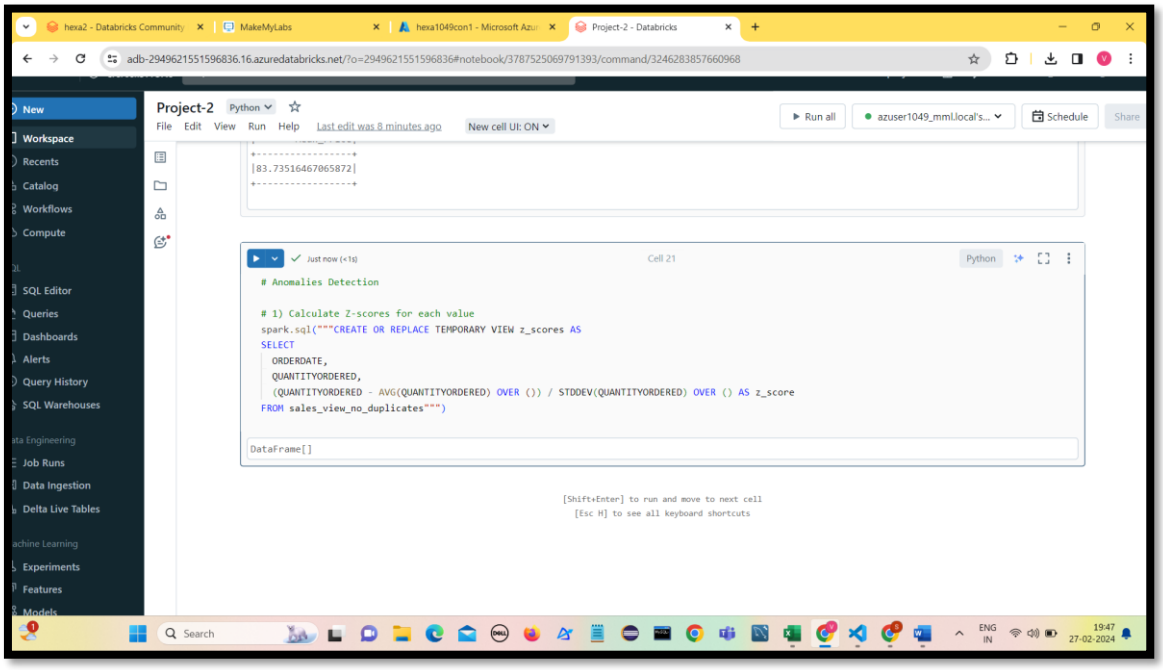
X is the data point,

μ is the mean, and

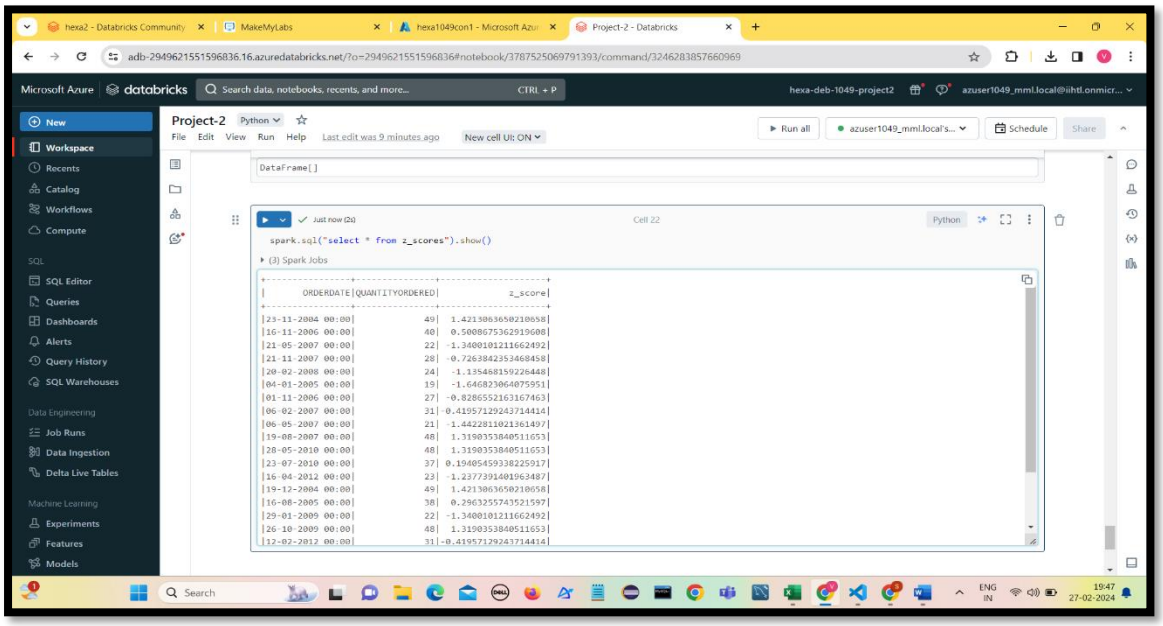
σ is the standard deviation.

AVG(value) OVER () calculates the mean value of the entire dataset.

STDDEV(value) OVER () calculates the standard deviation of the entire dataset.



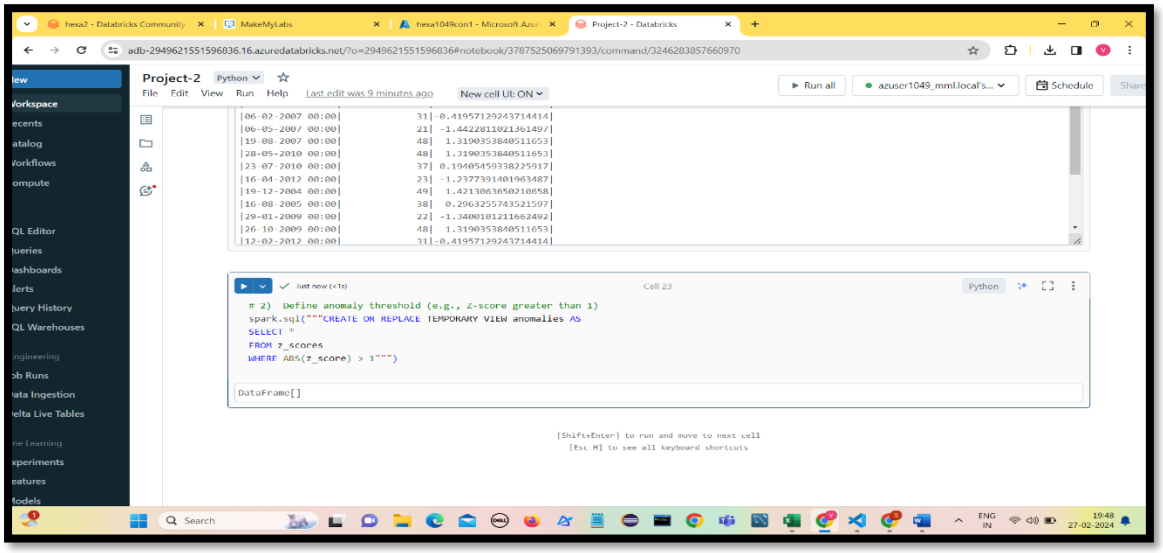
2) Display of z-scores



3) Define Anomaly Threshold

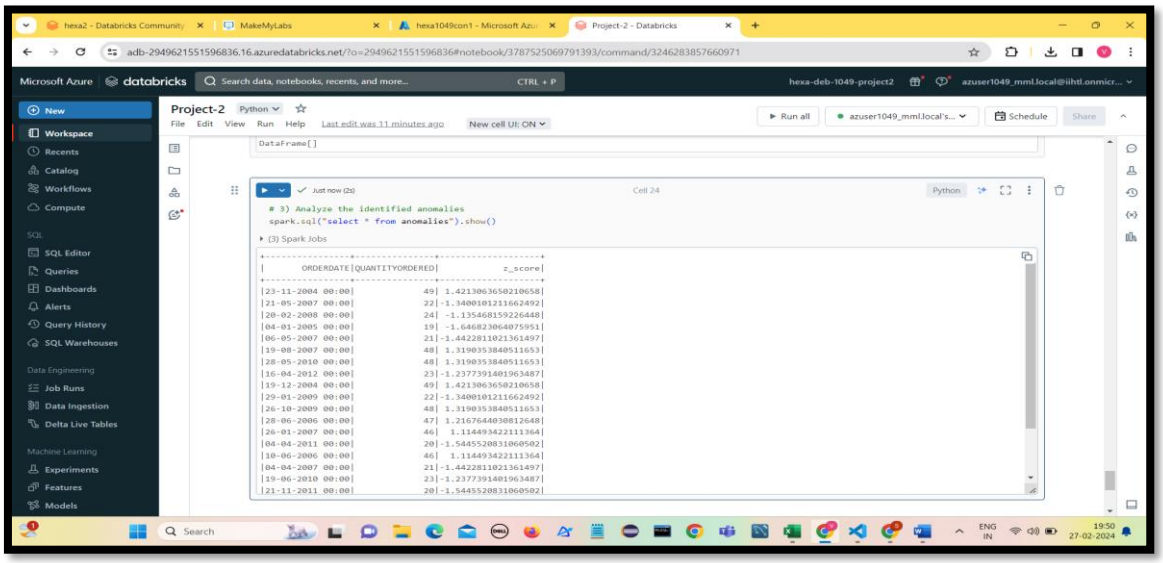
- An anomaly threshold is a predefined value that determines when a data point is considered an anomaly.
- Common thresholds include Z-scores greater than a certain number of standard deviations from the mean.

- We define the threshold as ABS(z_score) > 1 to identify anomalies that are more than 1 standard deviations away from the mean.



4) Analyze Anomalies

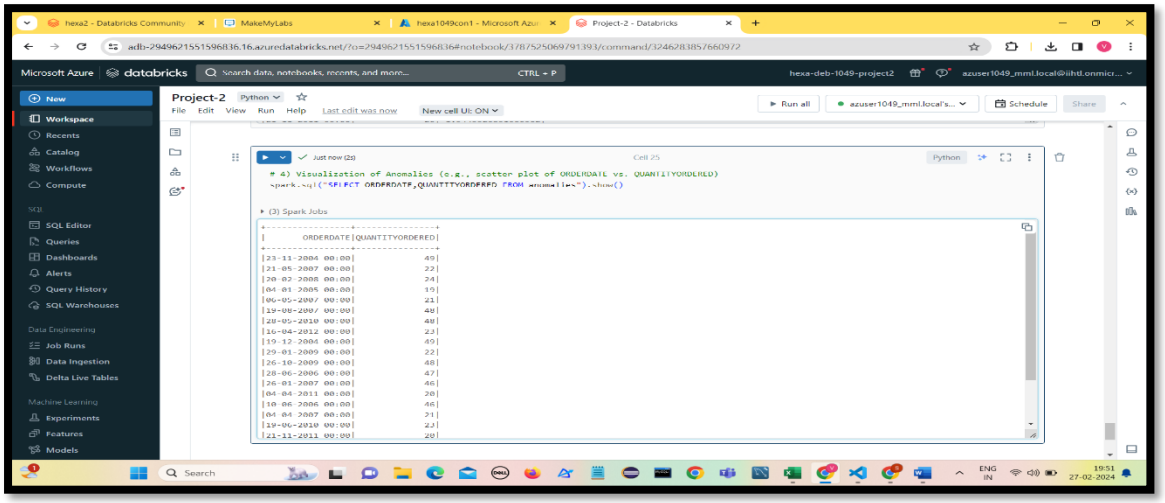
Analyzing the identified anomalies involves examining their timestamps, associated values, and potential causes. It's essential to investigate anomalies further to determine whether they are genuine outliers or errors in the data.

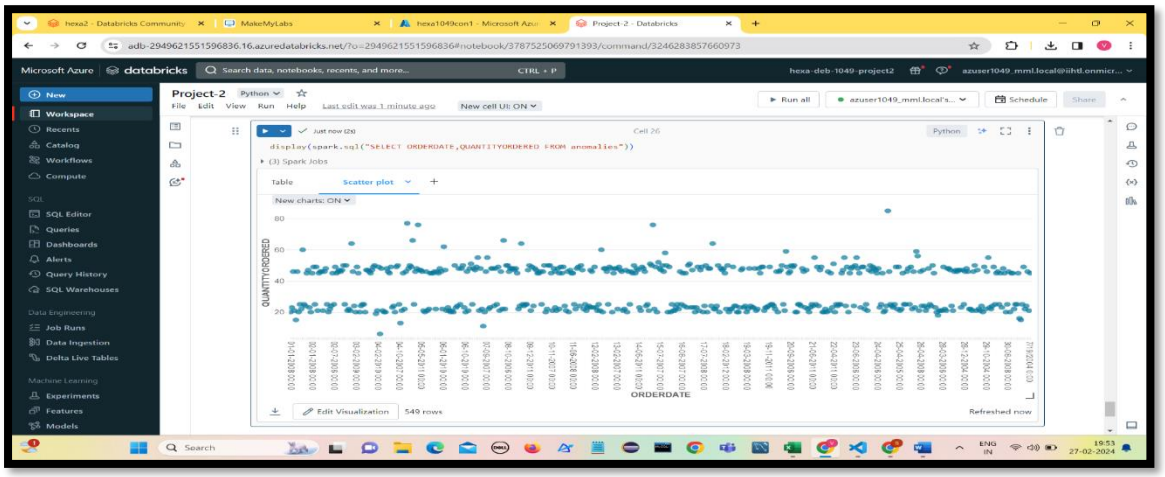


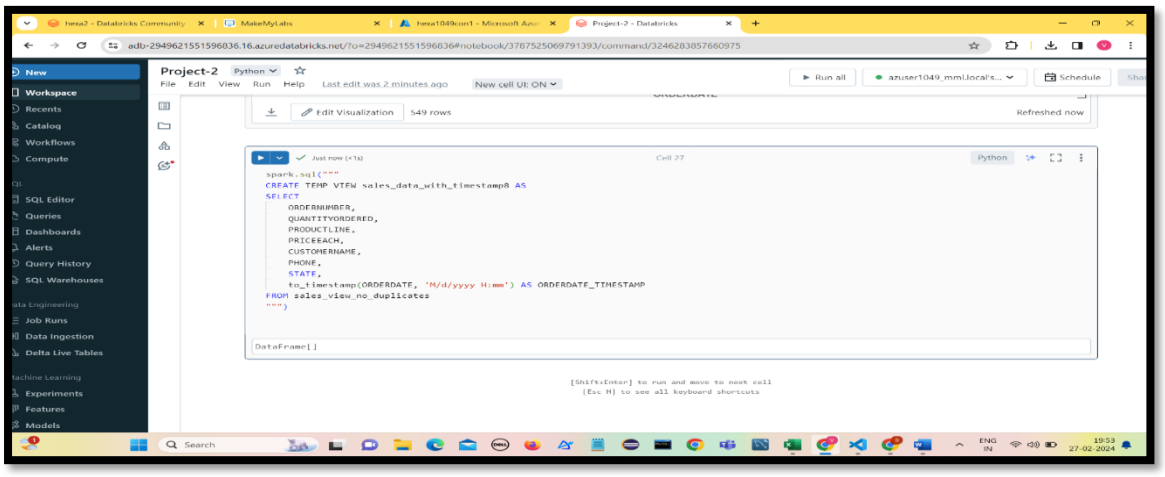
5) Visualize Anomalies

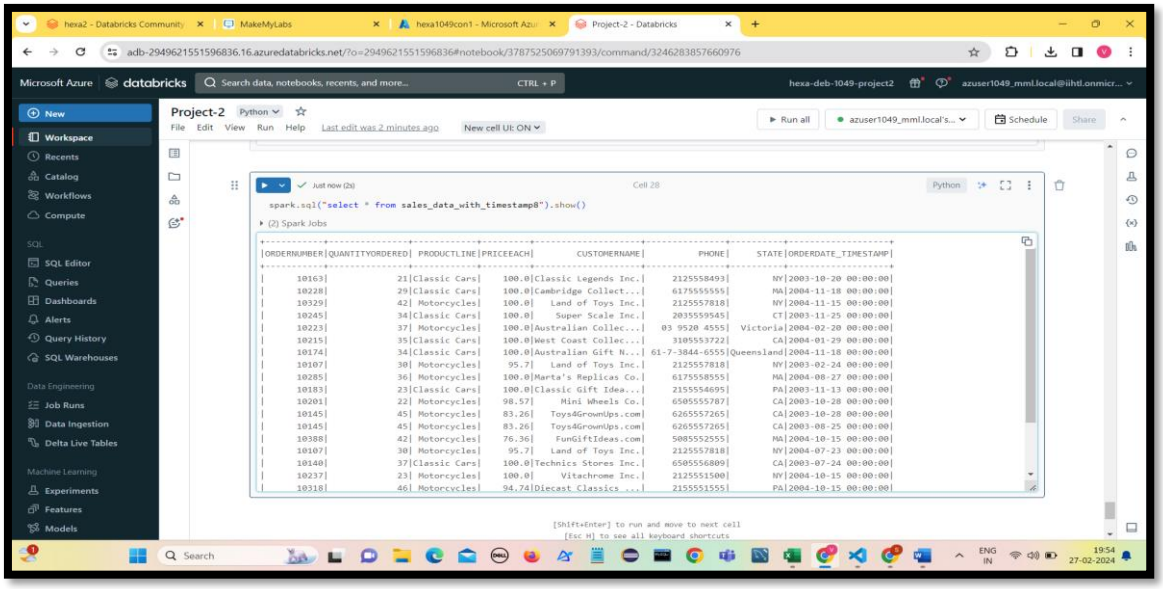
Visualizing anomalies can provide insights into their distribution and relationship with other variables in the dataset. This step can be helpful for gaining a deeper understanding of the anomalies.

Visualize using Scatter plot:







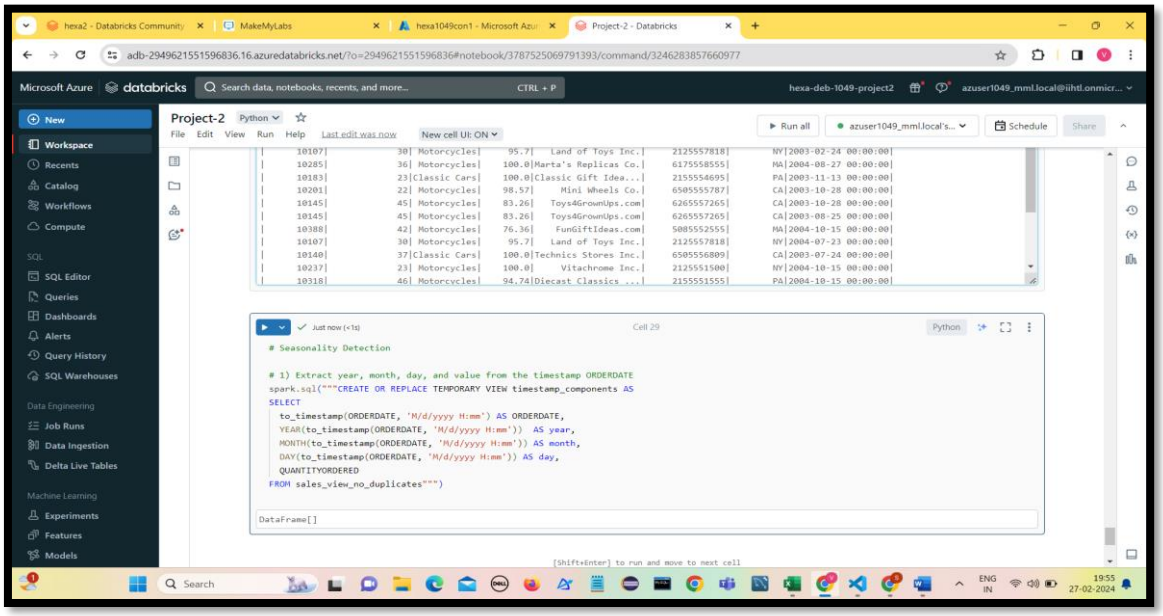


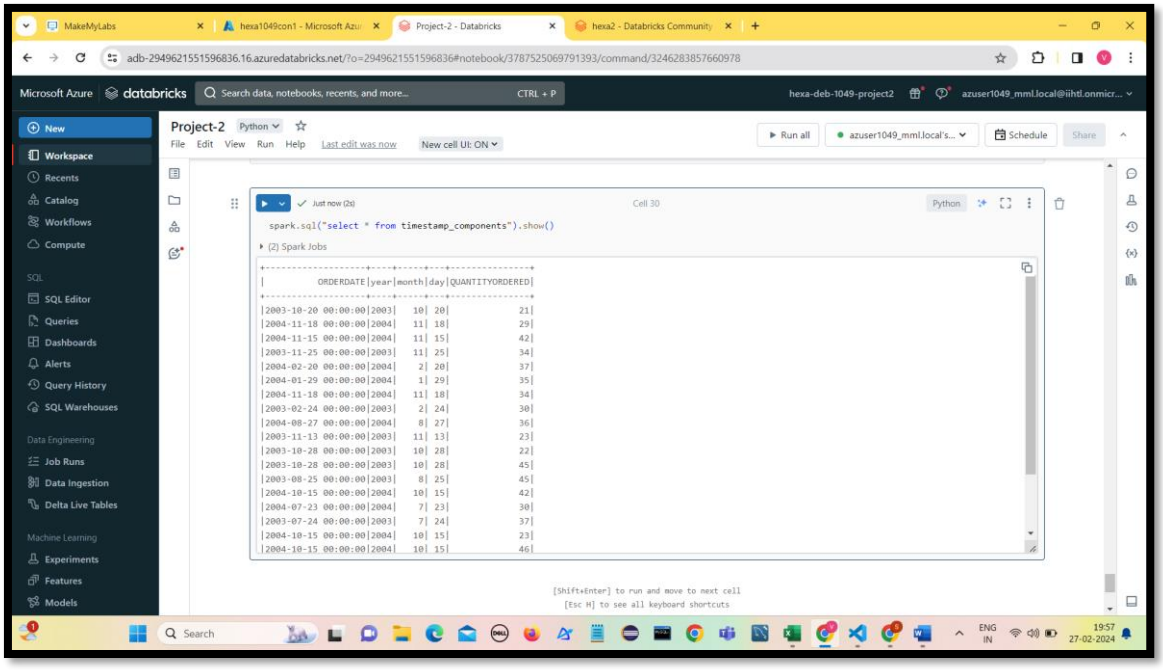
Seasonality Detection:

Seasonality refers to the repetitive and predictable patterns or fluctuations in a time series dataset that occur at regular intervals over time. These patterns typically repeat over specific periods, such as days, weeks, months, or years, and are often influenced by external factors such as weather, holidays, or economic cycles.

1) Extract Timestamp Components:

Extract year, month, day, and value from the timestamp ORDERDATE.

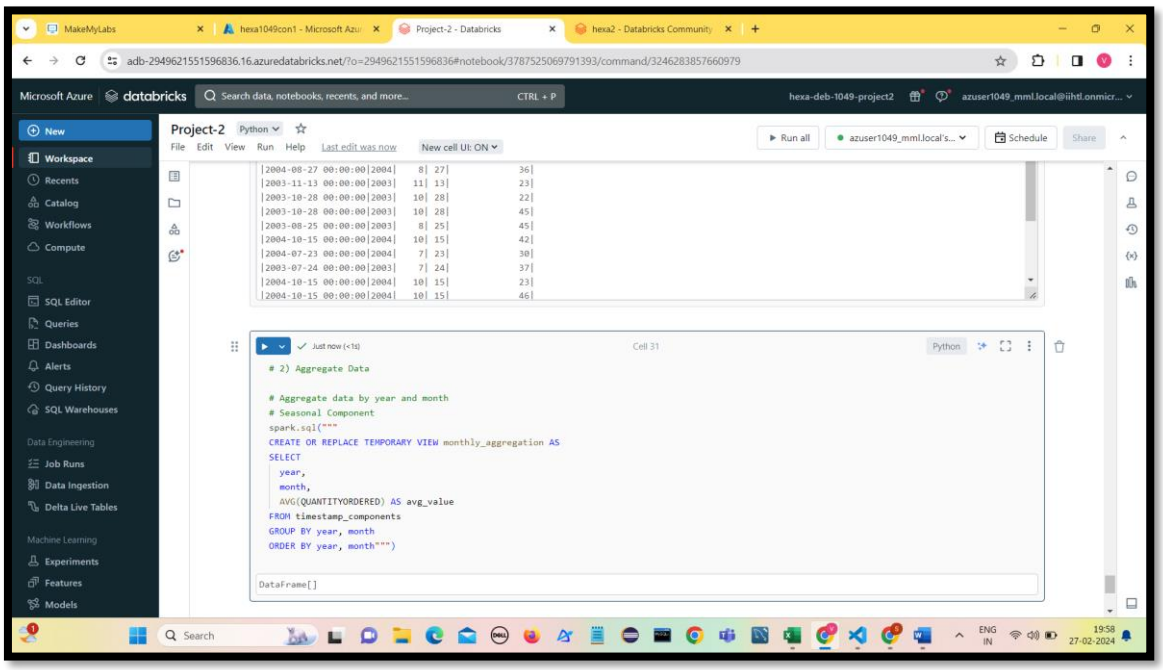




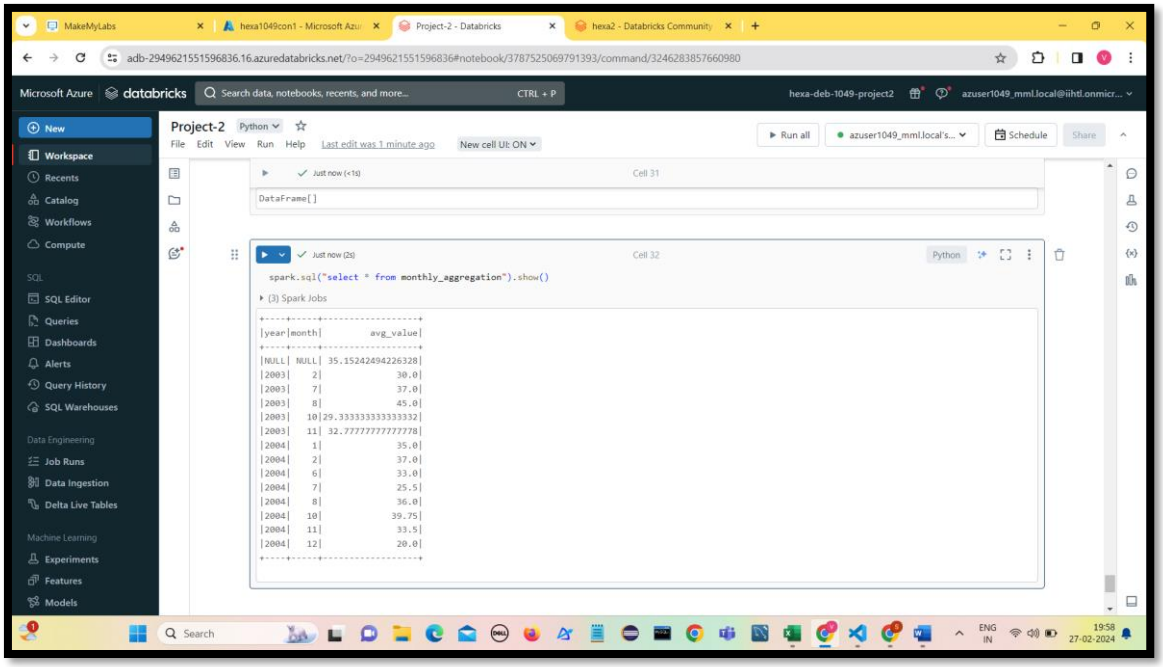
In this step, we extract relevant components such as year, month, day, and value from the timestamp column. These components will be used to analyze the seasonal patterns in the data.

2) Aggregate Data

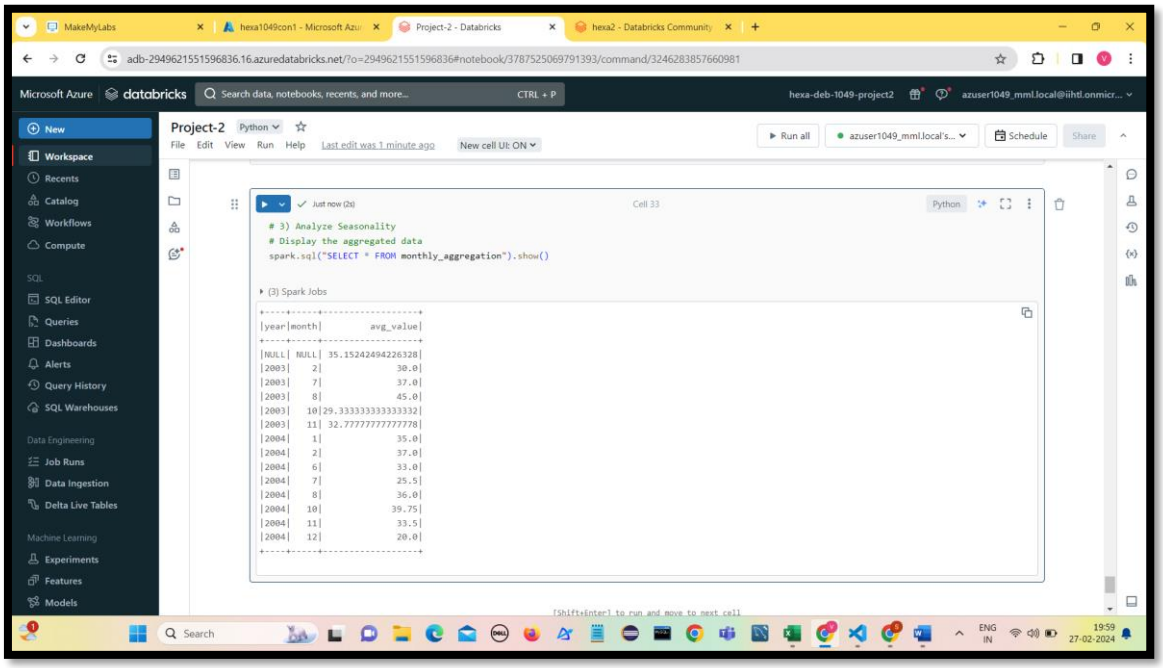
Aggregate data by year and month



In this step, we aggregate the data by year and month, calculating the average value for each month. This aggregated data will help us identify seasonal patterns over time.

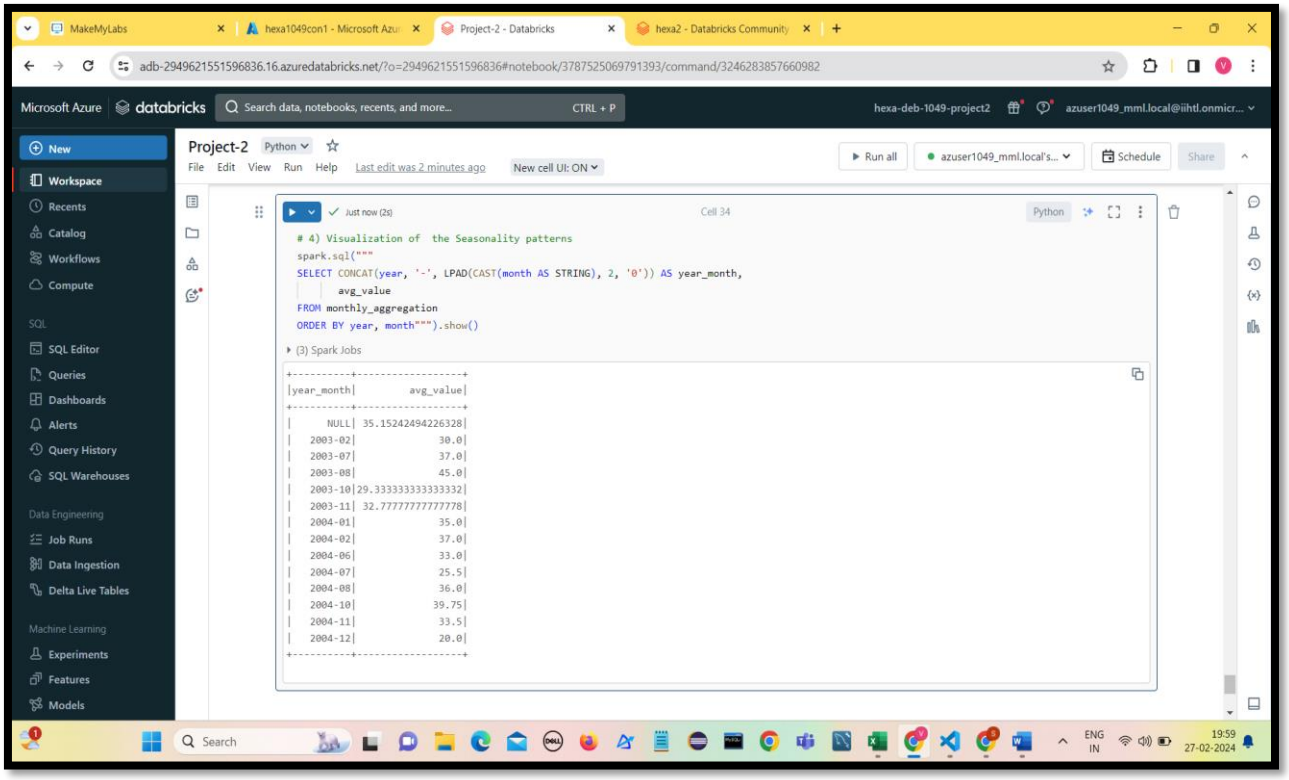


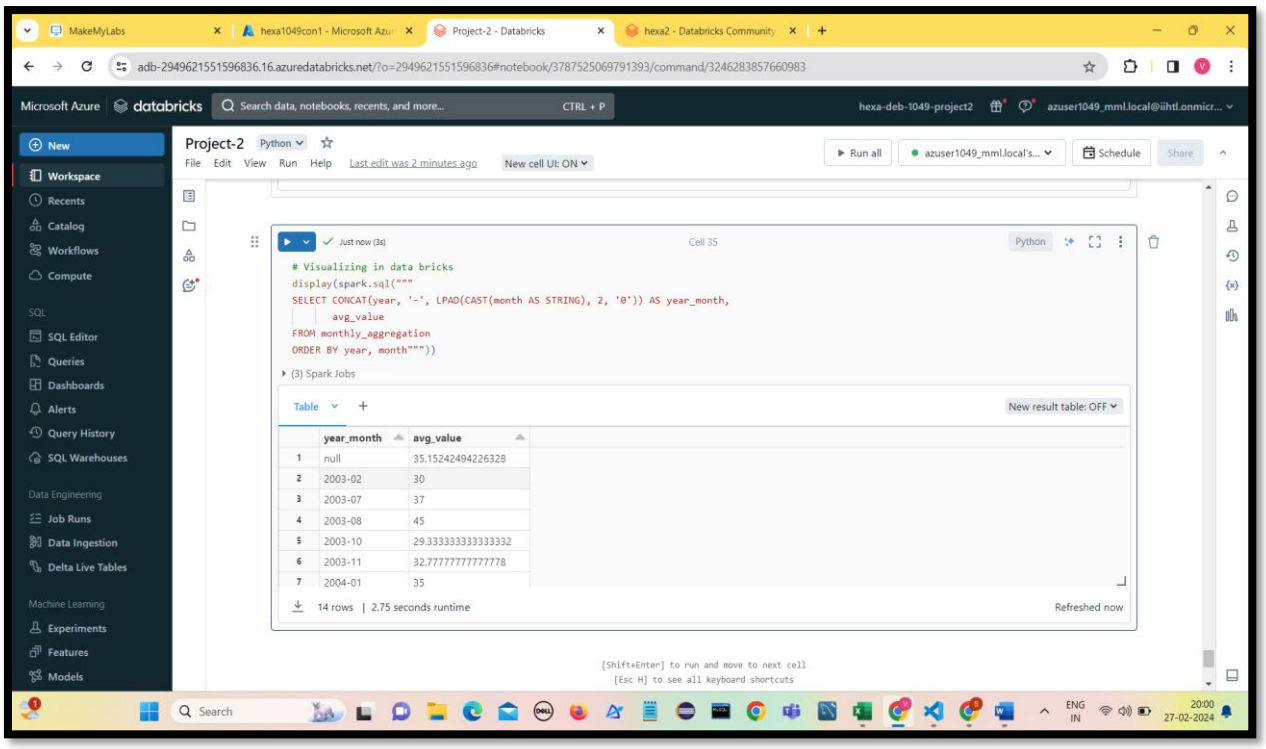
3) Analyse Seasonality:



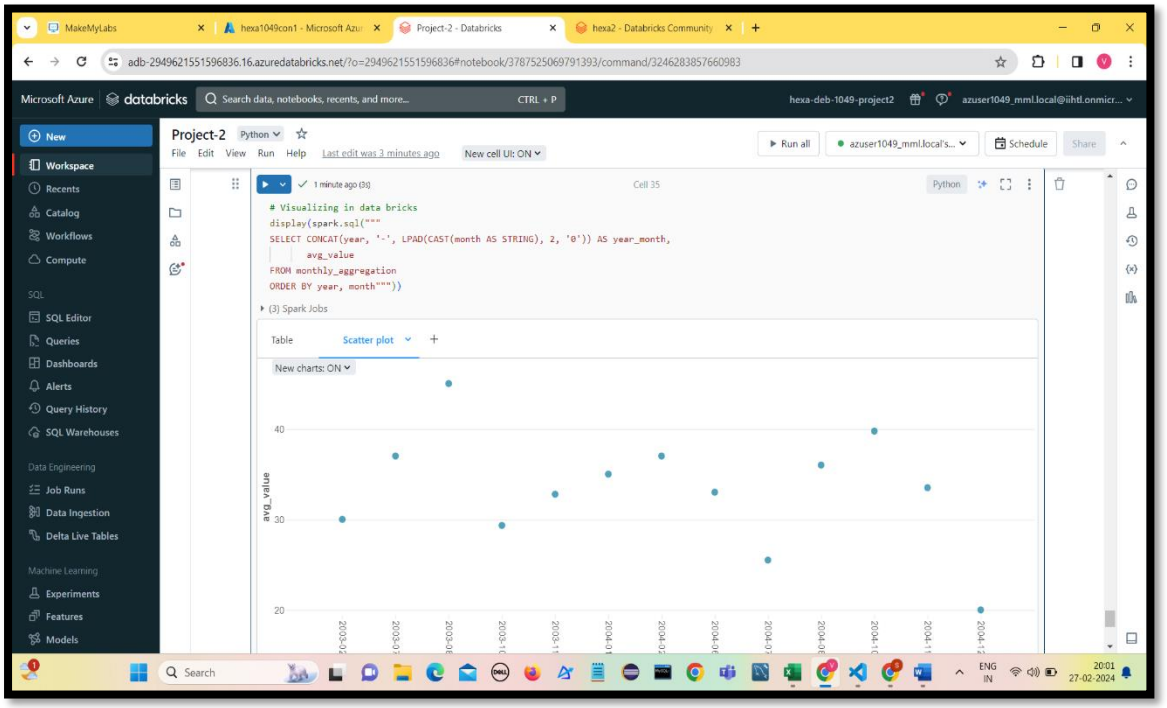
4) Visualize Seasonality

- In this step, we analyze the seasonality patterns by examining the aggregated data and visualizing it using line charts or other visualization tools.
- By observing the average value over time (e.g., months or years), we can identify any recurring patterns or trends indicating seasonality.

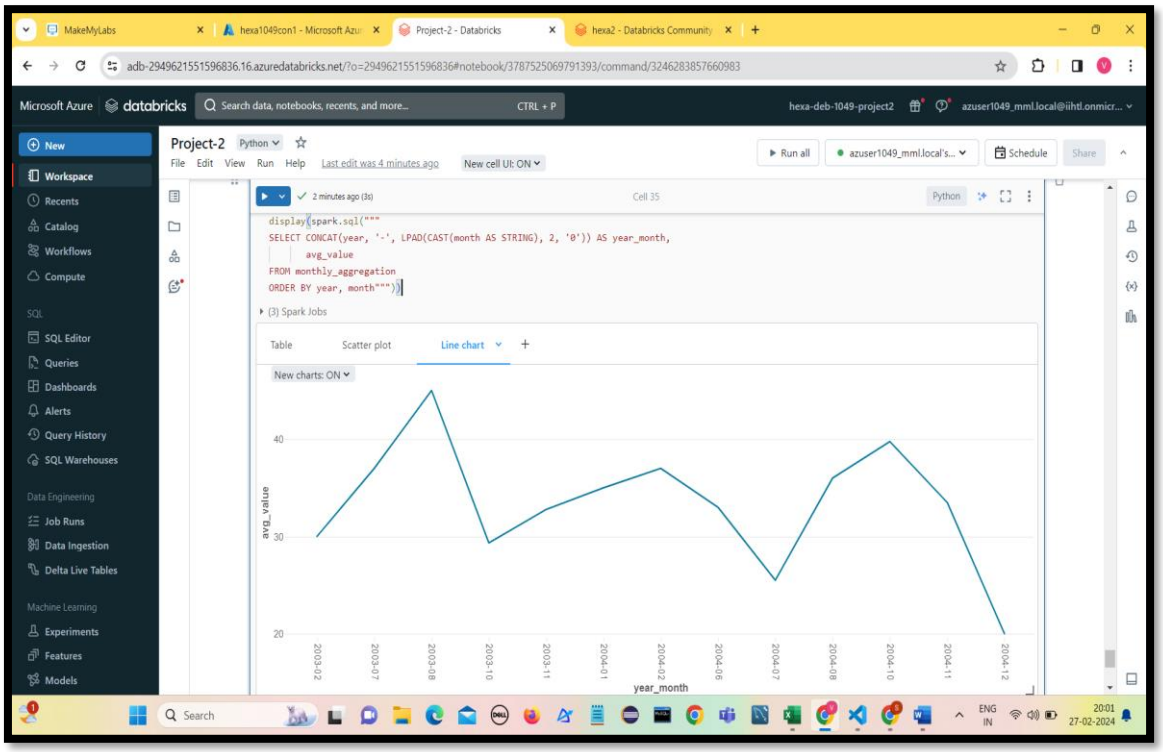

Project-2
Python
File Edit View Run Help Last edit was 2 minutes ago
New cell UI: ON
Run all
Schedule
Share
Cell 34
4) Visualization of the Seasonality patterns
spark.sql("""
SELECT CONCAT(year, '-', LPAD(CAST(month AS STRING), 2, '0')) AS year_month,
avg_value
FROM monthly_aggregation
ORDER BY year, month""").show()
(3) Spark Jobs
+----------+------------------+
|year_month| avg_value|
+----------+------------------+
NULL	35.15242494226328
2003-02	30.0
2003-07	37.0
2003-08	45.0
2003-10	29.333333333333332
2003-11	32.77777777777778
2004-01	35.0
2004-02	37.0
2004-06	33.0
2004-07	25.5
2004-08	36.0
2004-10	39.75
2004-11	33.5
2004-12	20.0
+----------+------------------+


Project-2
Python
File Edit View Run Help Last edit was 2 minutes ago
New cell UI: ON
Run all
Schedule
Share
Cell 35
Visualizing in data bricks
display(spark.sql("""
SELECT CONCAT(year, '-', LPAD(CAST(month AS STRING), 2, '0')) AS year_month,
avg_value
FROM monthly_aggregation
ORDER BY year, month"""))
(3) Spark Jobs
Table
New result table: OFF
year_month avg_value
1 null 35.15242494226328
2 2003-02 30
3 2003-07 37
4 2003-08 45
5 2003-10 29.333333333333332
6 2003-11 32.77777777777778
7 2004-01 35
14 rows | 2.75 seconds runtime
Refreshed now
[Shift+Enter] to run and move to next cell
[Esc H] to see all keyboard shortcuts

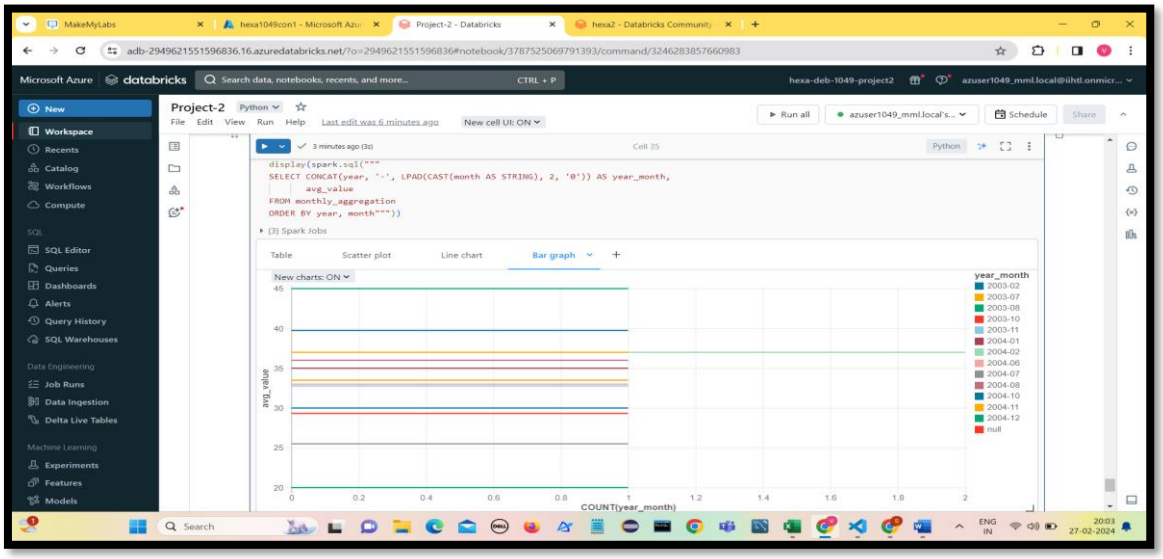
1) Scatter plot



2) Line Chart



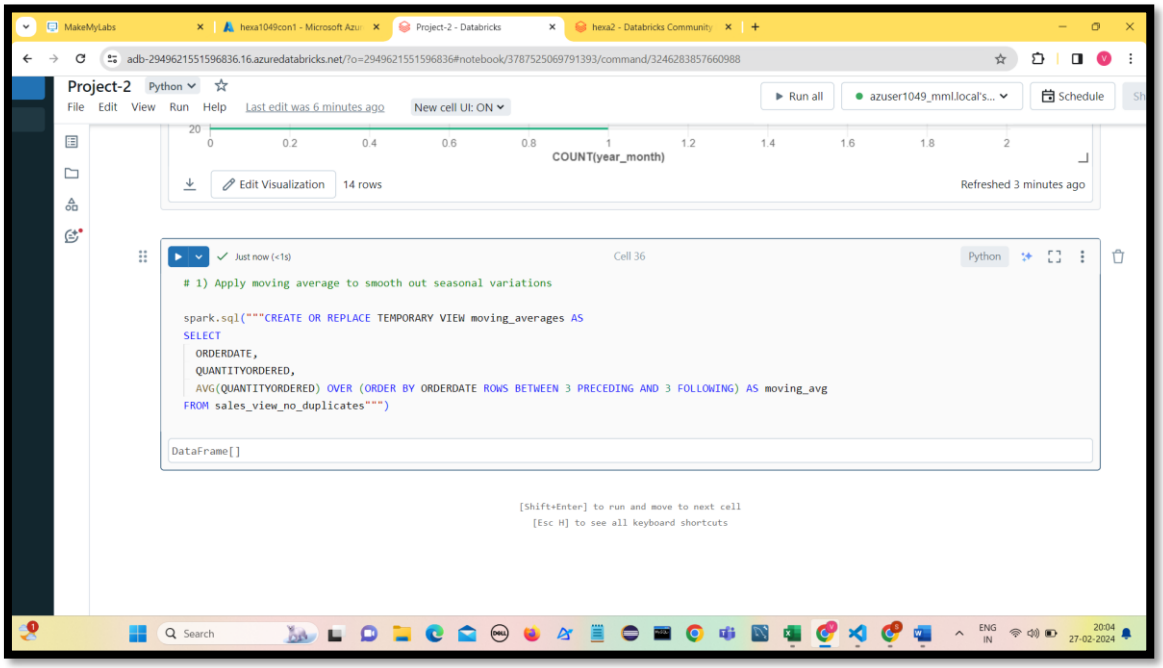
3) Bar chart



Moving Averages

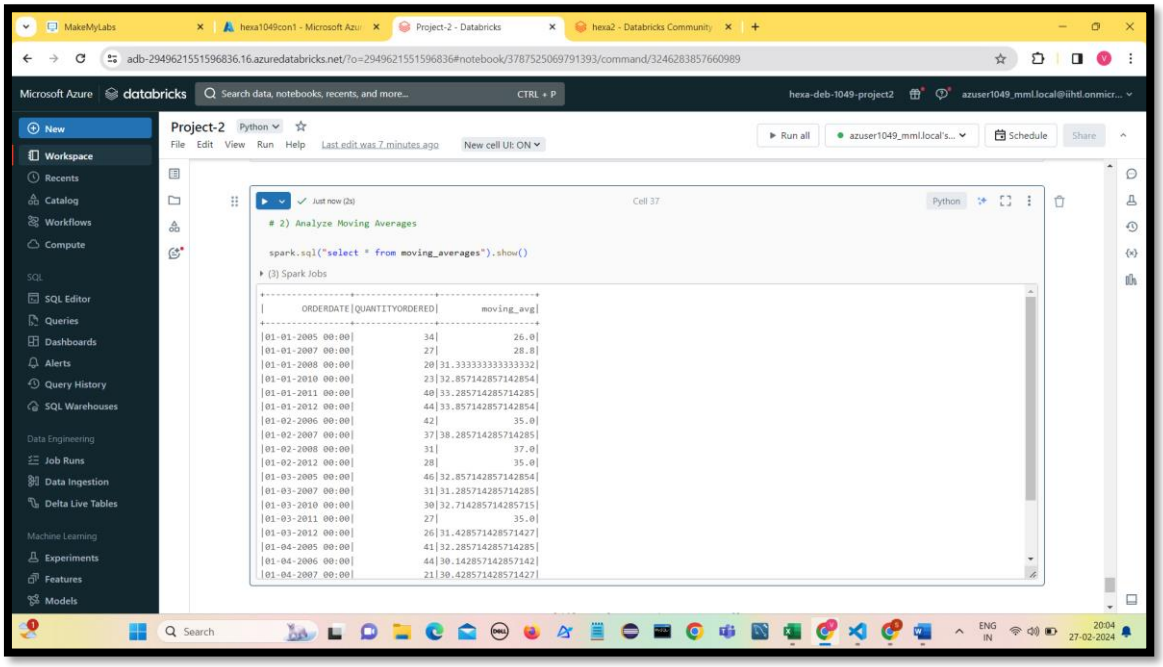
Moving averages are commonly used to smooth out fluctuations in time-series data and identify trends or patterns over a specific period. In the context of detecting seasonality using Spark SQL on Azure Databricks, we can apply moving averages to analyze the trend and smooth out the seasonal variations in the data.

1) Calculate moving averages



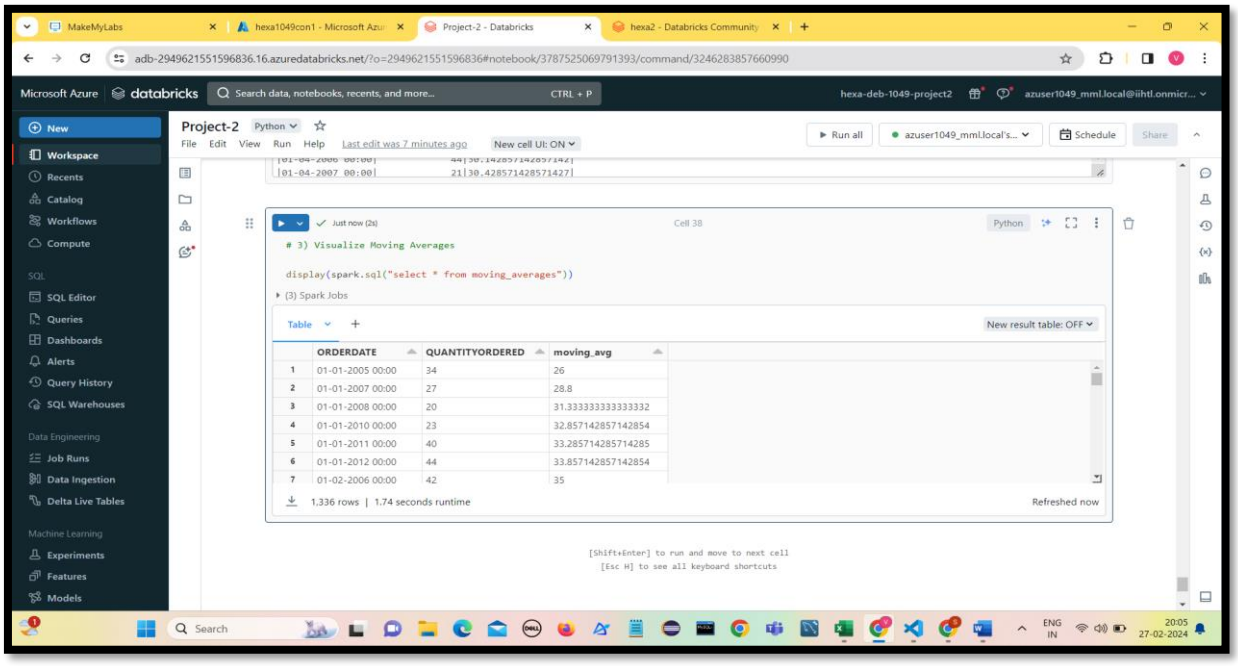
2) Analyze Moving Averages

Once you have calculated the moving averages, you can analyze them to identify trends and smooth out seasonal variations in the data.

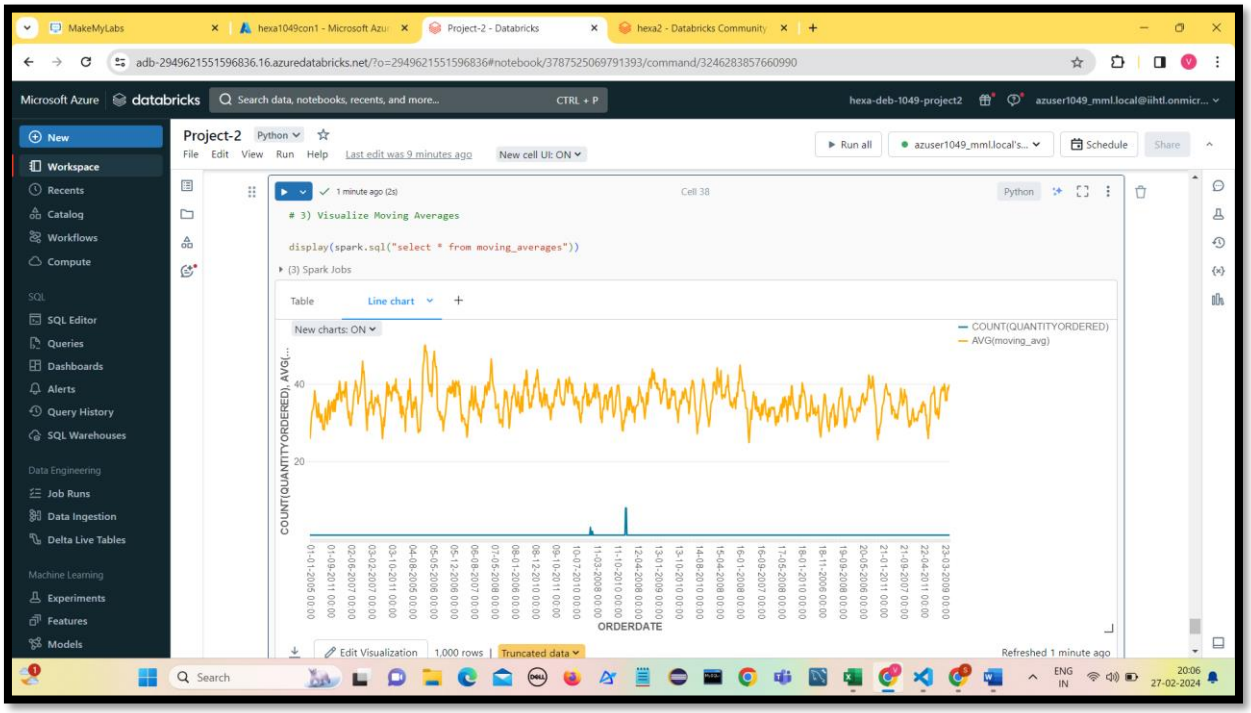


3) Visualize moving averages

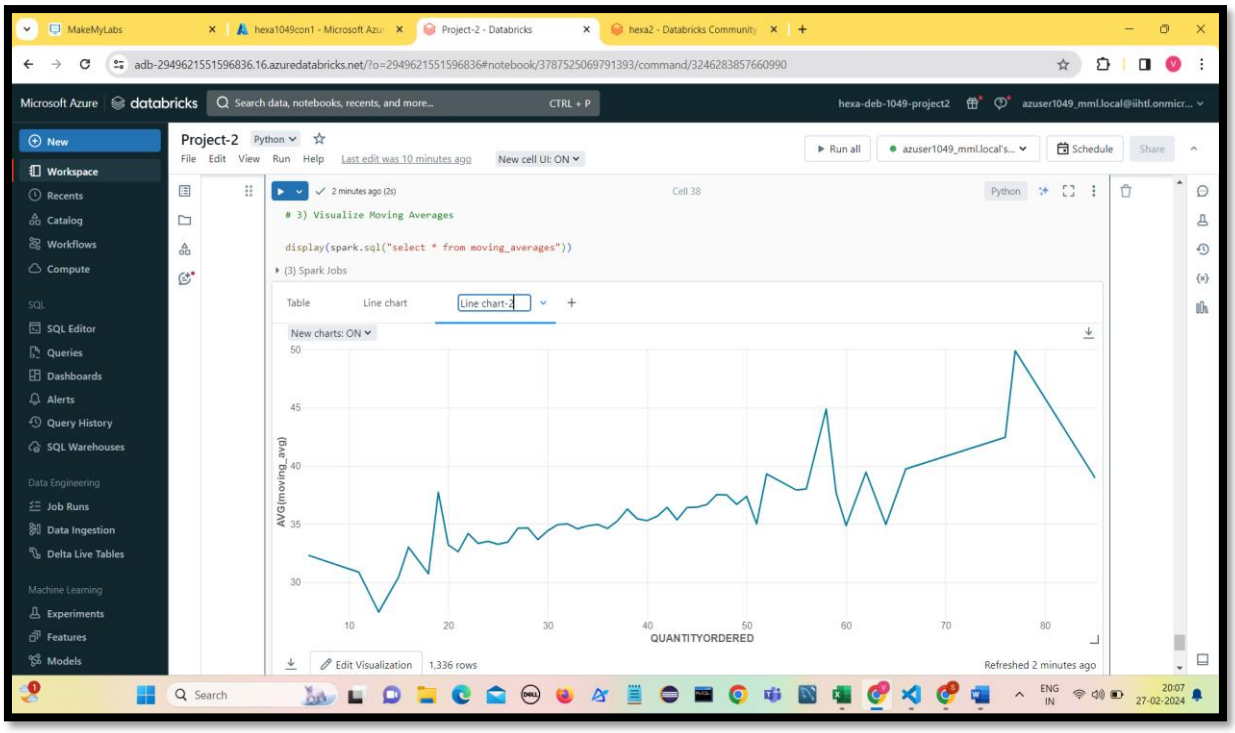
Visualization can provide a clearer understanding of the trend and seasonal variations in the data after applying moving averages.



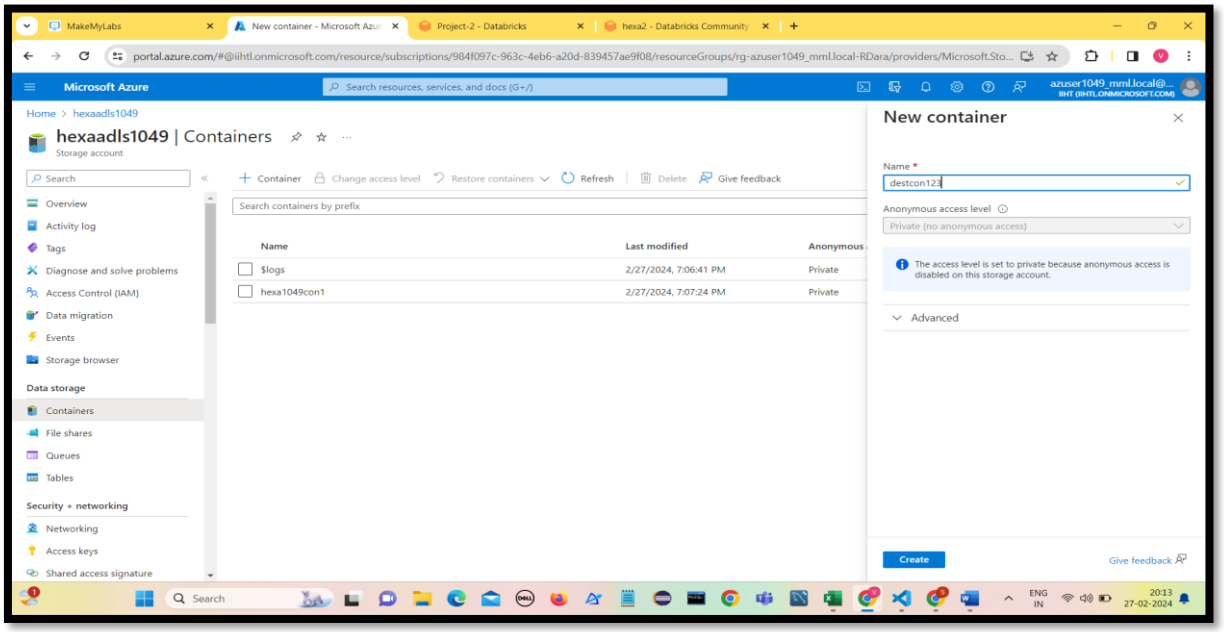
1) Line Chart



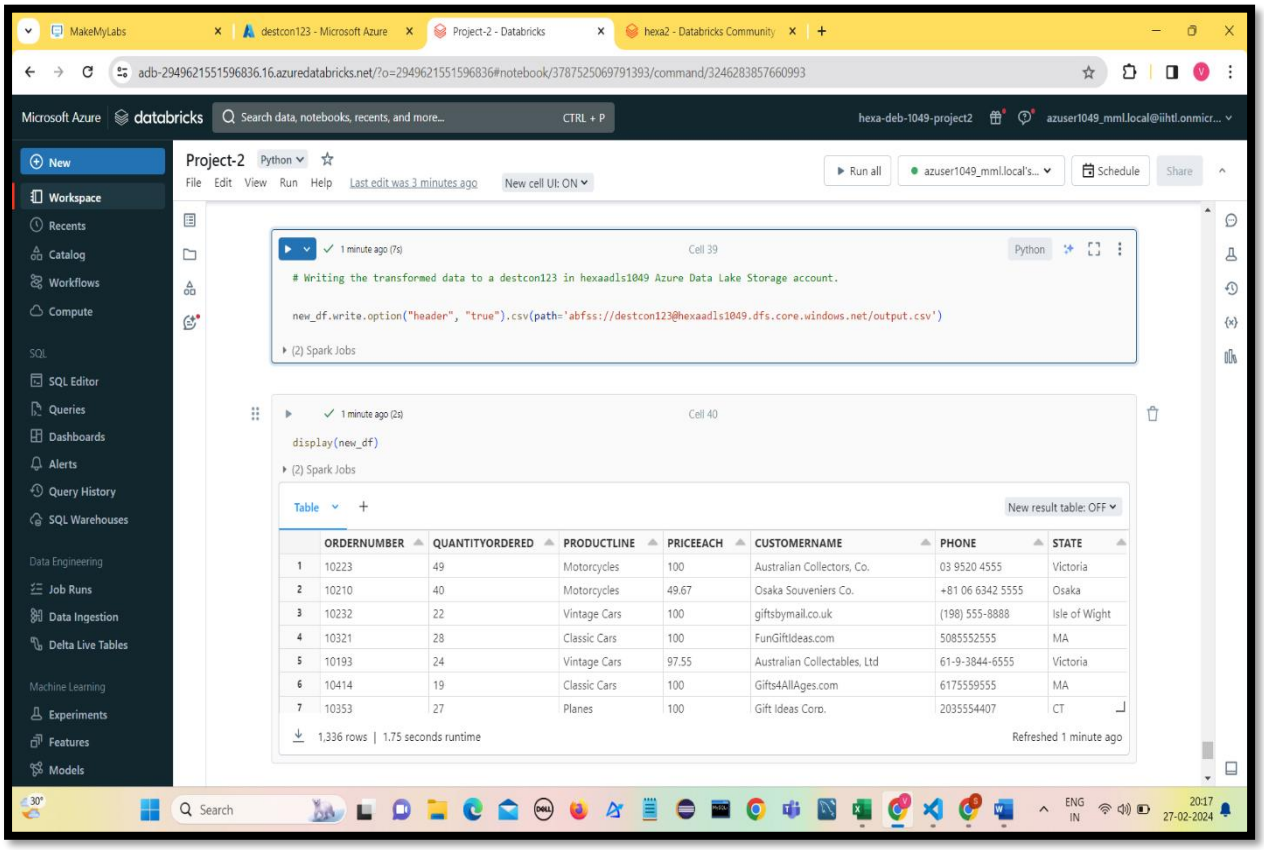
2) Line chart-2



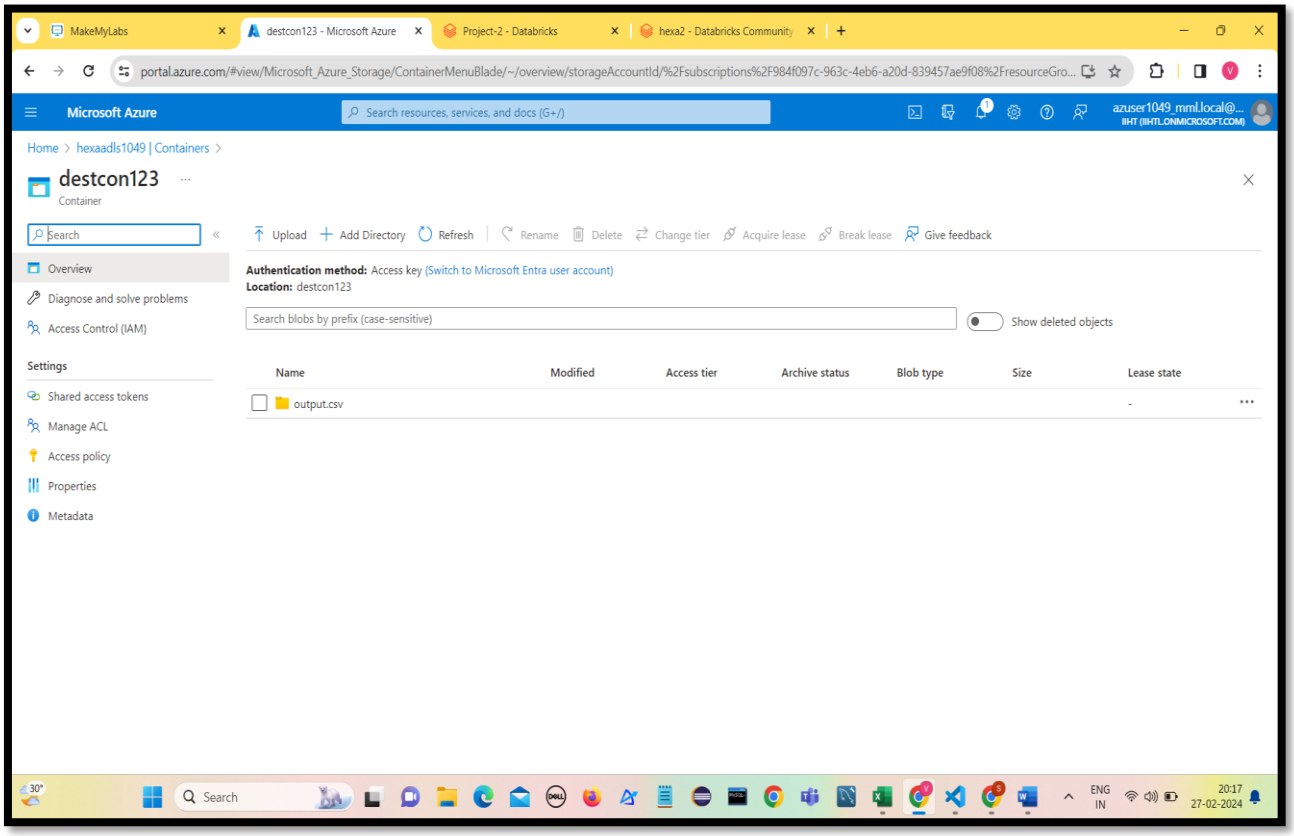
Create a destination container to write the transformed data.

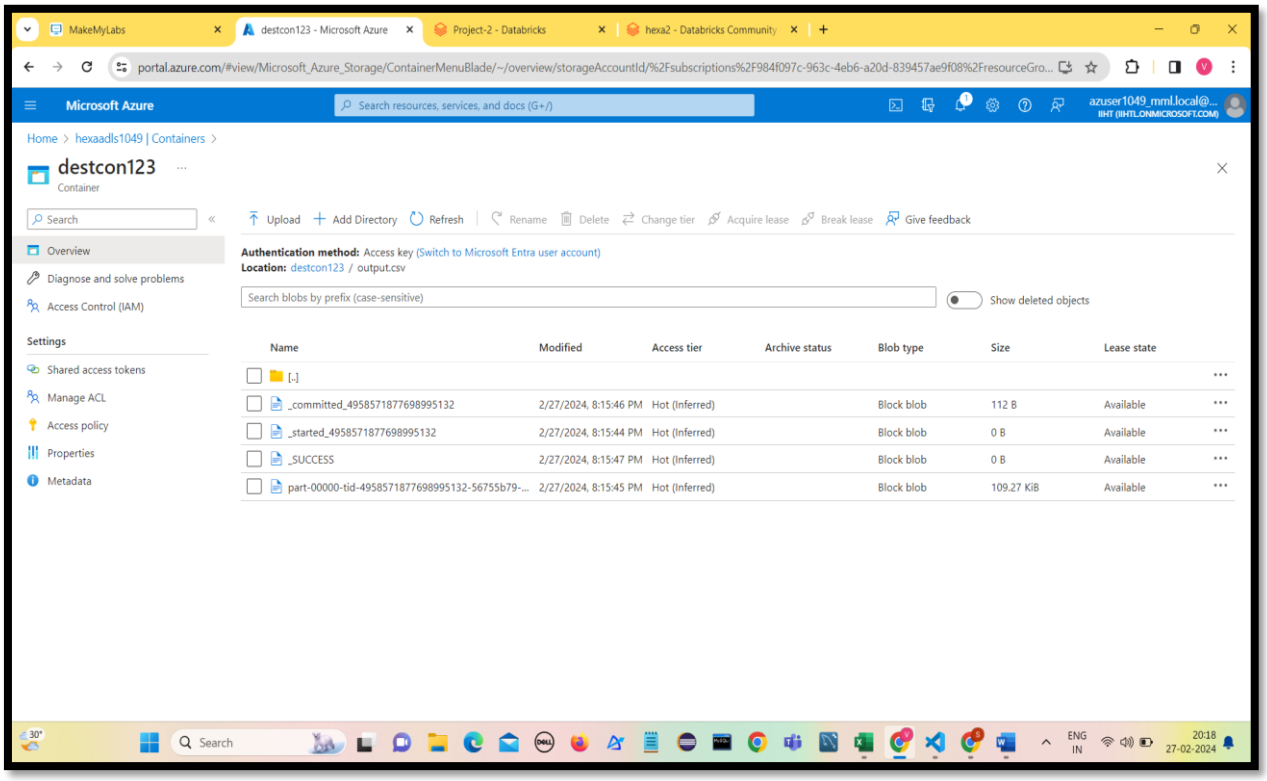


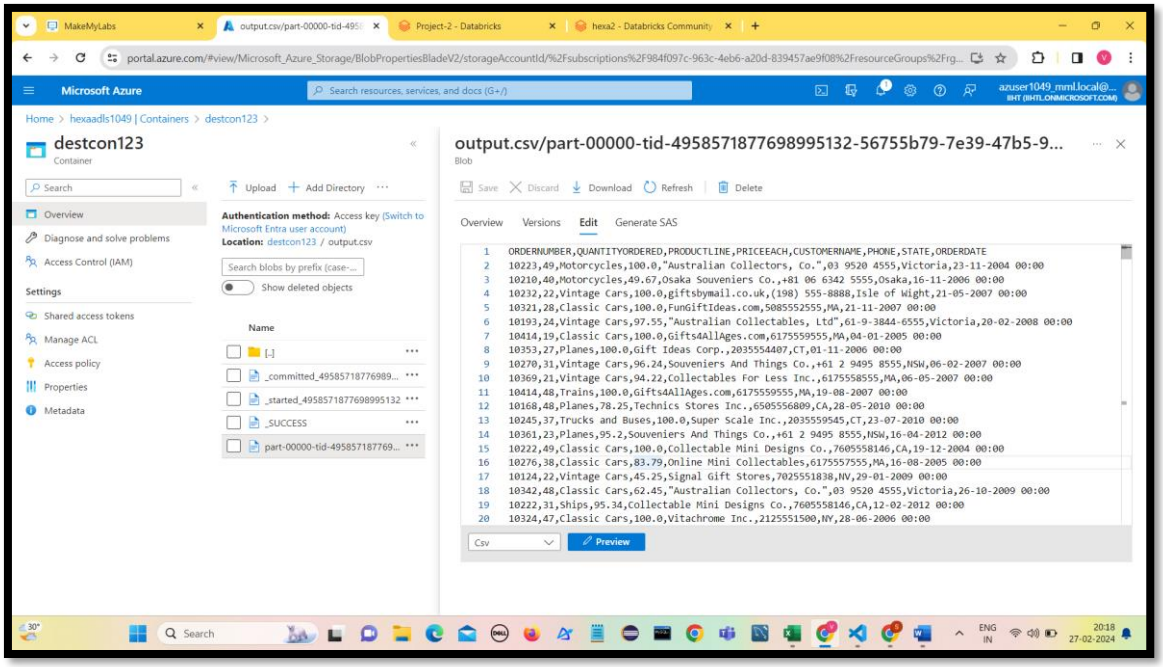
Write the transformed data to destcon123 in hexaadls1049 Azure Data Lake Storage



Writing into output.csv file







Conclusion:

In conclusion, the project focused on leveraging Spark SQL on Azure Databricks for data analysis. The project demonstrated the effectiveness of using Spark SQL on Azure Databricks for data analysis tasks, particularly in detecting seasonality and identifying anomalies in time series data. The insights gained from this project can inform strategic decision-making and operational planning for various domains, including retail, finance, healthcare, and more.

Vuthur Sriganga

Hexaware Data Engineering Batch-1

vuthursriganga@gmail.com