Praxis Business School

Bangalore

A logo with text on it

Description automatically generated

A PROJECT REPORT ON

***“Olympics Data Analysis: Microsoft Azure End-to-End Pipeline Project”***

**A report submitted in partial fulfilment of the requirements for the**

**Capstone Academic Project**

**Submitted by**

**Tejesh Kumar**  **H23011**

**Darshana Patil**  **H23003**

**Nikita Sargam H23008**

**Under the guidance of,**  
**Atanu Ghosh**

**atanu@praxis.ac.in**

**SUMMARY**

**Our project aims to create an end-to-end data pipeline using Microsoft Azure services for Olympics data analytics. The Olympics are not only a celebration of athletic prowess but also a goldmine of data waiting to be explored. By developing this pipeline, we intend to streamline the collection, cleaning, and transformation of Olympics data from diverse sources, ensuring its quality for analysis. Leveraging Azure's powerful analytics tools, including machine learning and statistical analysis, we aim to extract actionable insights such as athlete performance trends and predictors of medal outcomes.**

**These insights will be stored in Azure data repositories and visualized using tools like Power BI, providing stakeholders with intuitive dashboards for informed decision-making. By empowering Olympic committees, coaches, athletes, sponsors, and media outlets with data-driven insights, our project seeks to enhance strategic planning, resource allocation, and overall Olympic experiences. Ultimately, our endeavour underscores the transformative potential of data analytics in sports, contributing to the ongoing success and innovation of the Olympic movement.**

**Table of Contents**

* [**Table of Contents**](https://github.com/SherlockkOms/Tokyo-Olympics-Data-Analysis-Microsoft-Azure-End-to-End-Pipeline-Project?tab=readme-ov-file#table-of-contents)
* [**Project Overview**](https://github.com/SherlockkOms/Tokyo-Olympics-Data-Analysis-Microsoft-Azure-End-to-End-Pipeline-Project?tab=readme-ov-file#project-overview)
* [**Project Architecture**](https://github.com/SherlockkOms/Tokyo-Olympics-Data-Analysis-Microsoft-Azure-End-to-End-Pipeline-Project?tab=readme-ov-file#project-architecture)
* [**Technologies**](https://github.com/SherlockkOms/Tokyo-Olympics-Data-Analysis-Microsoft-Azure-End-to-End-Pipeline-Project?tab=readme-ov-file#technologies-and-services-used)
* [**Data Ingestion Process**](https://github.com/SherlockkOms/Tokyo-Olympics-Data-Analysis-Microsoft-Azure-End-to-End-Pipeline-Project?tab=readme-ov-file#data-ingestion-process)
* [**Data Transformation Process**](https://github.com/SherlockkOms/Tokyo-Olympics-Data-Analysis-Microsoft-Azure-End-to-End-Pipeline-Project?tab=readme-ov-file#data-transformation-process)
  + [**Creating the app:**](https://github.com/SherlockkOms/Tokyo-Olympics-Data-Analysis-Microsoft-Azure-End-to-End-Pipeline-Project?tab=readme-ov-file#creating-the-app)
  + [**Granting app access to the storage account:**](https://github.com/SherlockkOms/Tokyo-Olympics-Data-Analysis-Microsoft-Azure-End-to-End-Pipeline-Project?tab=readme-ov-file#granting-app-access-to-the-storage-account)
* Azure Data Bricks Processing
* [**Creation of the Database using Azure Synapse Analytics**](https://github.com/SherlockkOms/Tokyo-Olympics-Data-Analysis-Microsoft-Azure-End-to-End-Pipeline-Project?tab=readme-ov-file#creation-of-the-database-using-azure-synapse-analytics)
* [**Generating a report in Power-BI**](https://github.com/SherlockkOms/Tokyo-Olympics-Data-Analysis-Microsoft-Azure-End-to-End-Pipeline-Project?tab=readme-ov-file#generating-a-report-in-powerbi)
* [**Conclusion**](https://github.com/SherlockkOms/Tokyo-Olympics-Data-Analysis-Microsoft-Azure-End-to-End-Pipeline-Project?tab=readme-ov-file#conclusion)

**Project Overview**

**The goal of this project is to develop an end-to-end data pipeline to ingest, transform, analyse, and visualize historic data from the Tokyo Olympics.**

**Initially, data is ingested into an Azure Data Lake Gen2 storage account using Azure Data Factory pipelines. Following ingestion, the data is processed and curated using Databricks Notebooks. Azure Synapse Analytics is then deployed to create a SQL database from the curated data, which is subsequently integrated into Power-BI for sophisticated analytics and visualizations.**

**Project Architecture**

**The architecture of this data pipeline has been designed to be robust and scalable. It consists of the following Azure services:**

* **Azure Databricks for data transformation.**
* **Azure Data Factory for data ingestion.**
* **Azure Storage (Data Lake Gen2) for data storage.**
* **Azure Synapse Analytics for creating a SQL database.**
* **Power BI for data visualization.**

A diagram of data pipeline

Description automatically generated

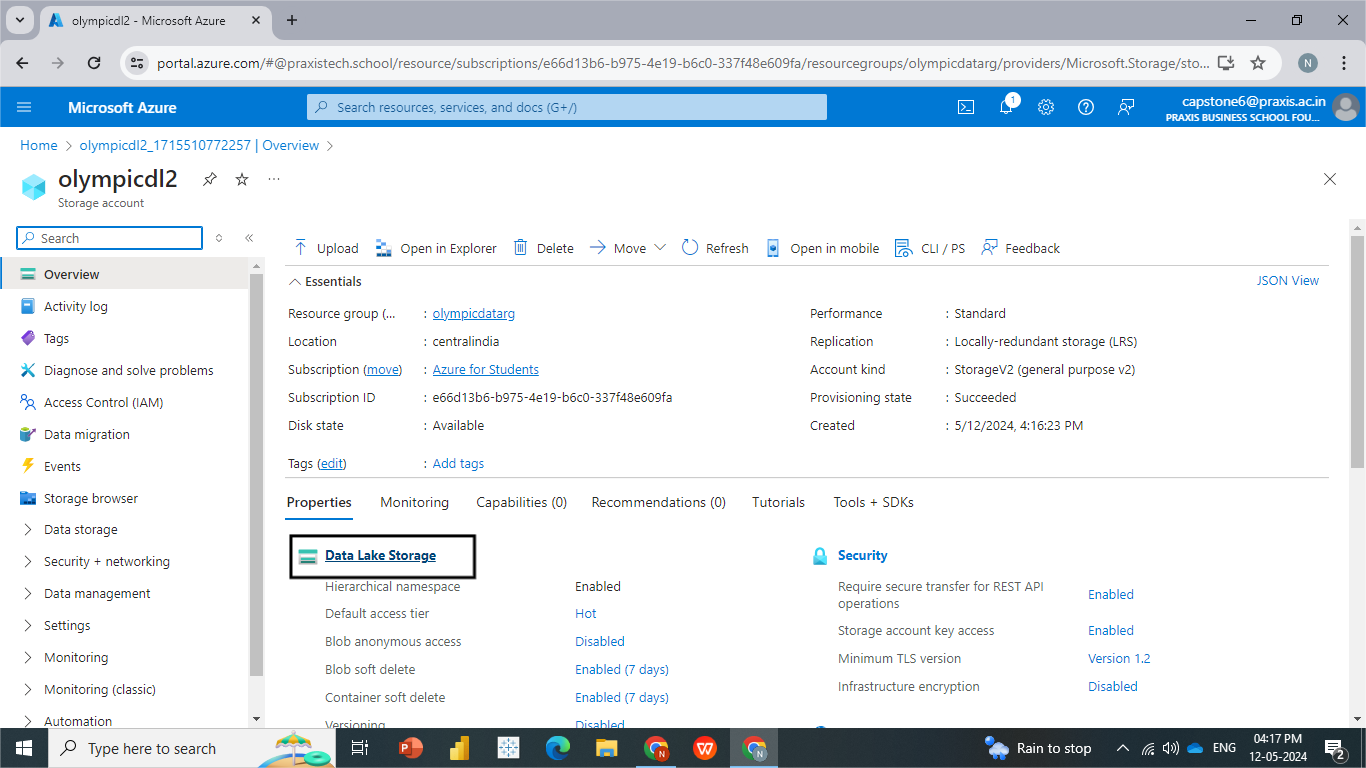
## **Technologies**

* **Azure Data Factory:** 
  + **It's a cloud-based data integration service that allows you to create, schedule, and orchestrate data workflows. It enables the movement and transformation of data from various data stores, both on-premises and in the cloud, for analytics, data integration, and intelligence applications.**
* **Databricks:** 
  + **An analytics platform optimized for the Microsoft Azure cloud services platform. It provides a collaborative Apache Spark-based analytics platform designed to process large amounts of data and integrate with various data sources and services in Azure.**
* **Azure Synapse Analytics:** 
  + **An integrated analytics service that combines big data and data warehousing. It offers a unified environment for preparing, managing, and serving data for immediate BI and machine learning needs, enabling end-to-end analytics solutions.**
* **Power BI:** 
  + **A business analytics service provided by Microsoft that provides interactive visualizations and business intelligence capabilities. With it, users can create reports and dashboards by connecting to a wide variety of data sources, simplifying data exploration, and sharing insights across an organization.**

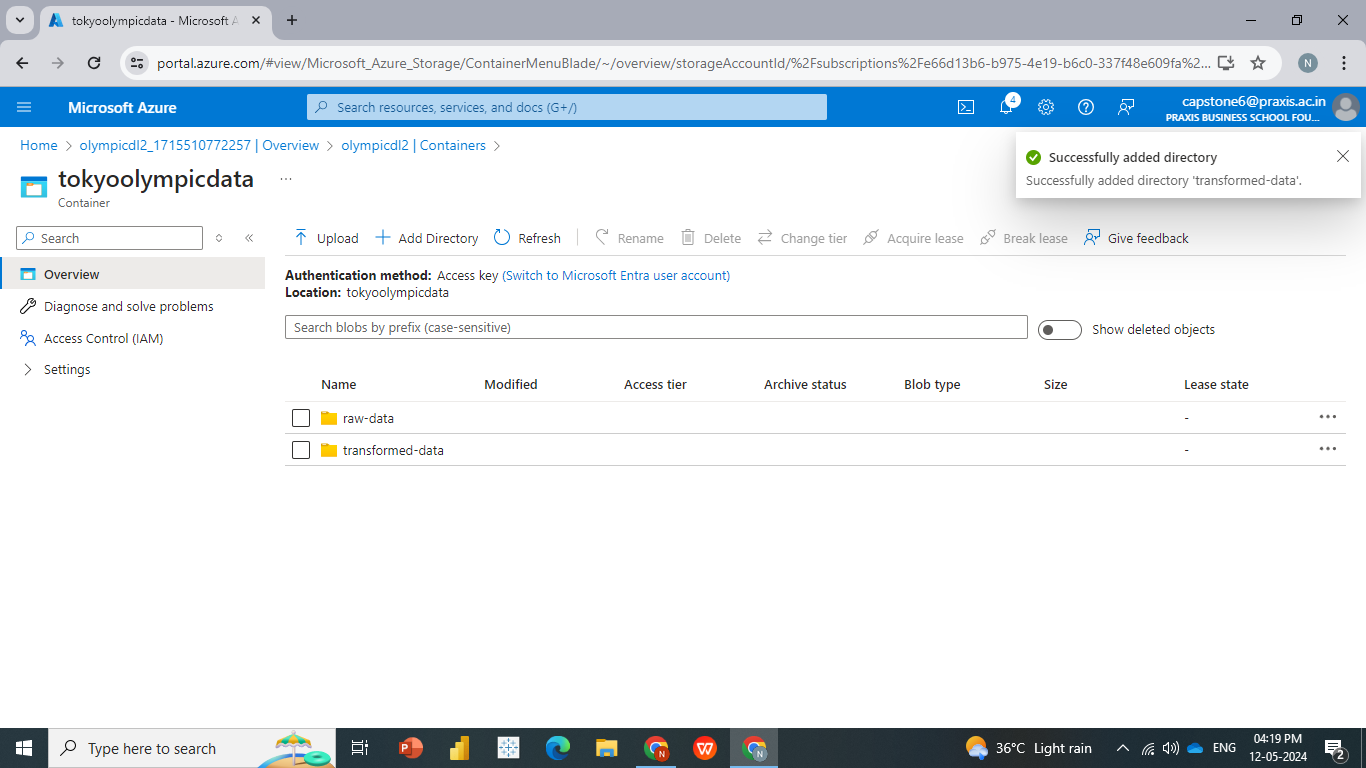
## **Data Ingestion Process**

**We employ ADF pipelines to ingest data into our Azure Data Lake Gen2 object storage. To do this we would first have to create a Linked Service to connect ADF to our Storage Account**

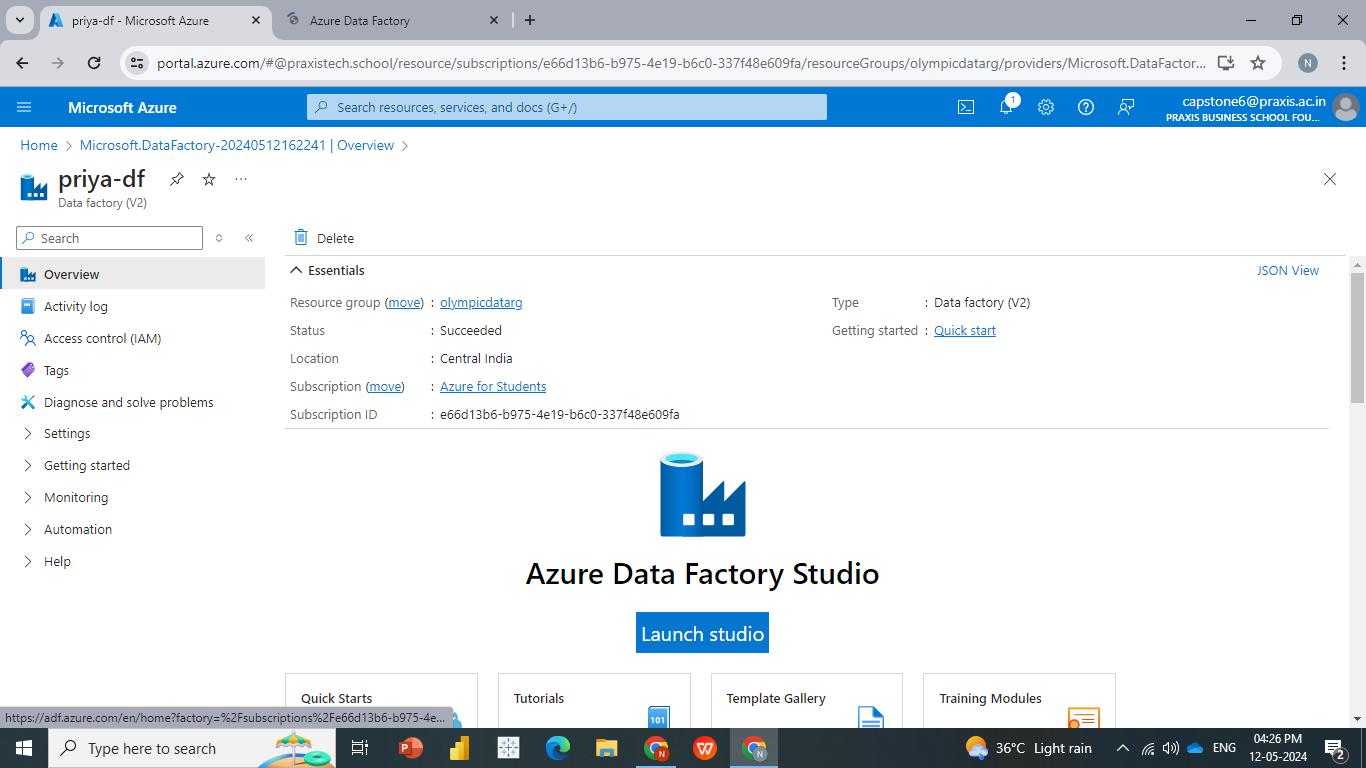
* **Login on to portal.azure.com**
* **Create a storage account and create a new resource group under which all instances get stored under one umbrella.**
* **Once Storage Account is created, then deployment of storage account takes place and once deployment is completed then you will find something like this as shown below.**



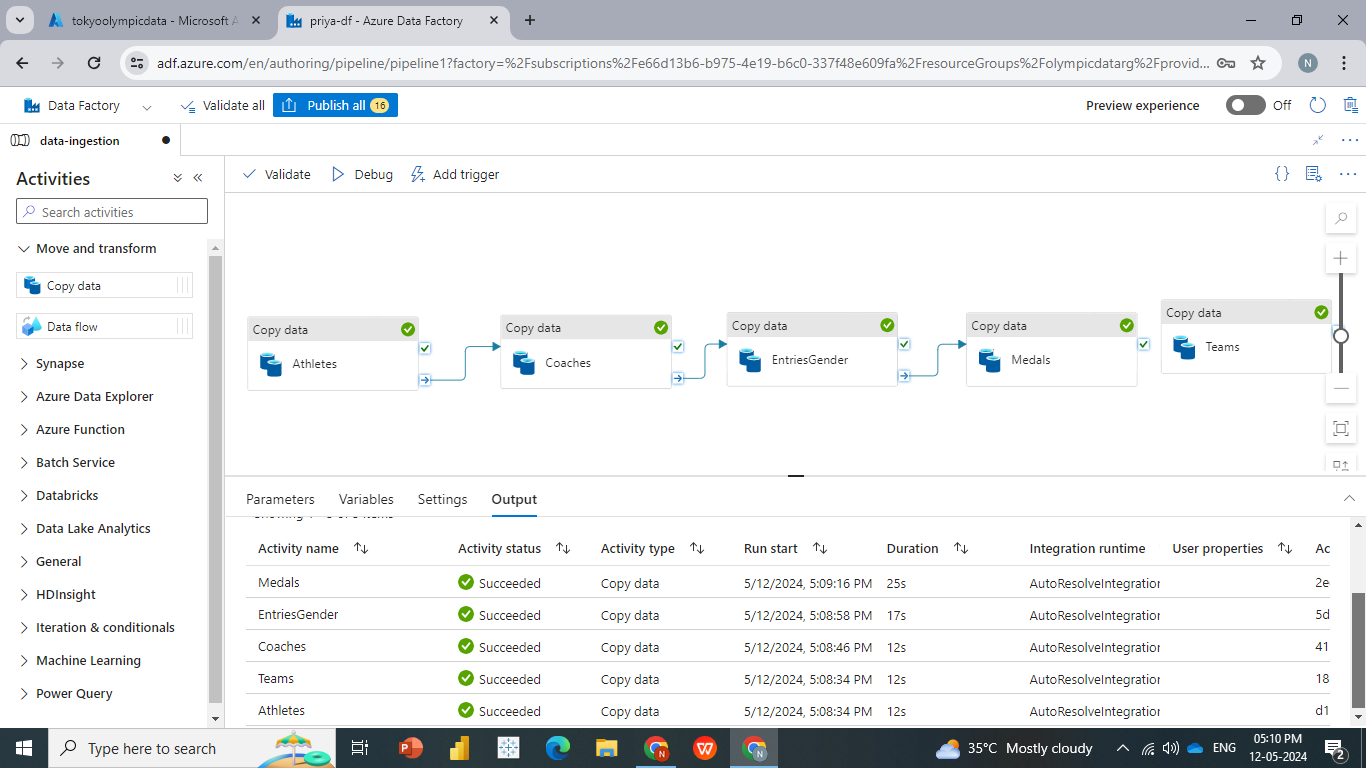
* **Now click on DATA LAKE STORAGE and create Container and give it a suitable name, where you can store data, as objects.**
* **After creation of container, click on add directories, we have created two directories one to store Raw data from data source and other to store Transformed data as shown below.**



* **Once we have established a connection, we can create an Ingestion Pipeline on Azure Data Factory to pull the data from online sources and save it to our storage container. Our online data source is primarily raw files available under the data tab on GitHub.**
* **Create a Data Factory and Launch Studio**
* **Review and create, once deployment is completed Launch Studio, it will redirect to Azure data factory panel, where u can build pipeline to extract data from various sources and upload it on to target location.**



* **Create a new pipeline and name it accordingly.**
* **From Move and Transform panel drag and drop Copy data into pipeline window, copy data will copy data from API/data-source to our location i.e., data storage.**
  + **Source:( where our data is stored)**
  + **Sink: (where we are supposed to load our data)**
* **Click on to source dataset -------> New Dataset ----> choose HTTP (Since we are extracting our data from GitHub repository and load onto Azure location.**
* **Create a New Linked Service, as we require raw data, choose Raw option from GitHub, copy URL, and paste it in field of Base URL, and click on Create.**
* **Load data on to data storage, choose Sink ----> New ----> Azure Data Lake Gen 2 ----> select format (Delimited Text) ---- > Create New Linked Service ---- > choose storage account --- > create --- > select path where u are supposed to store data( select Raw-data ) folder and give file name as athlete.csv ---- > validate the pipeline and debug.**
* **Now repeat the procedure for all other datasets (medals, entries-gender, coaches, teams)**
* **As we can see that our entire pipeline has been validated**



* **Here we can see that all our data has been loaded into Raw-data folder successfully, here we have completed data ingestion and loading of raw data.**

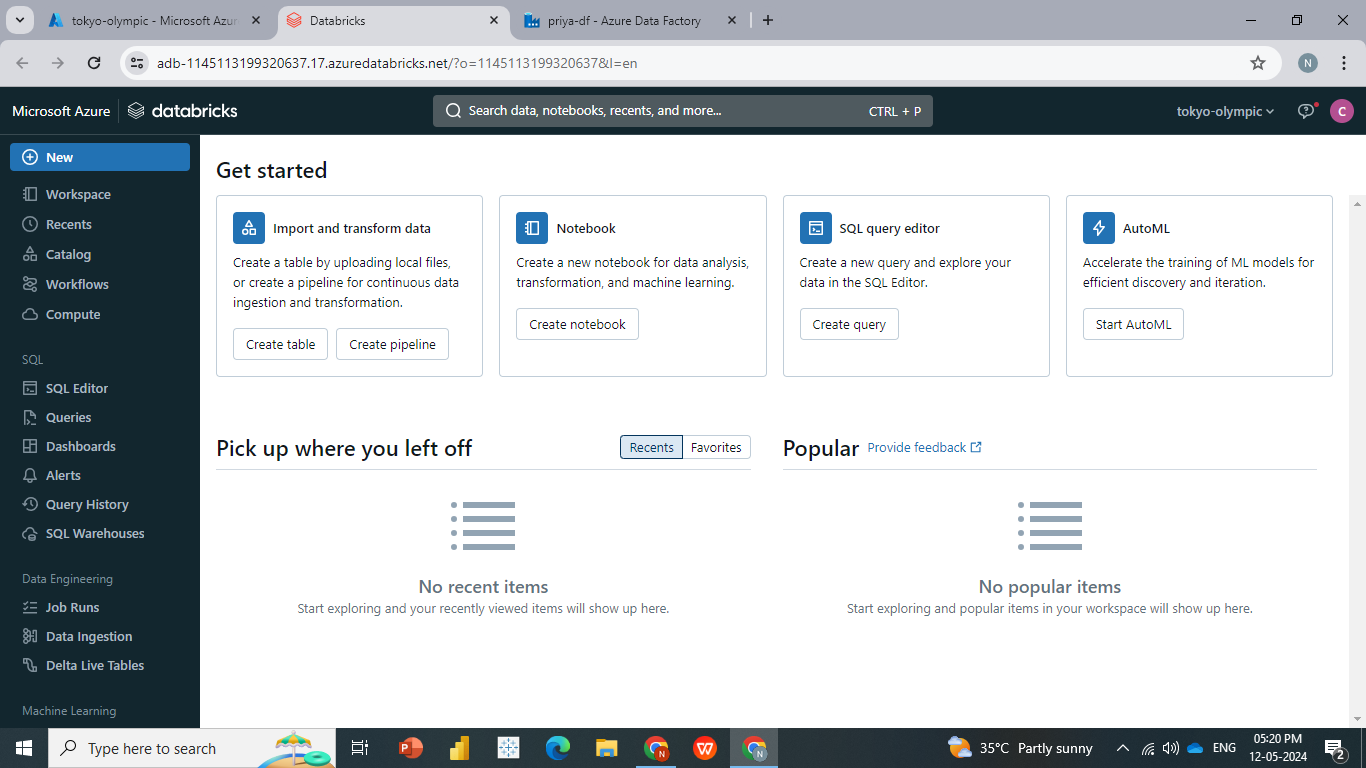
A screenshot of a computer

Description automatically generated

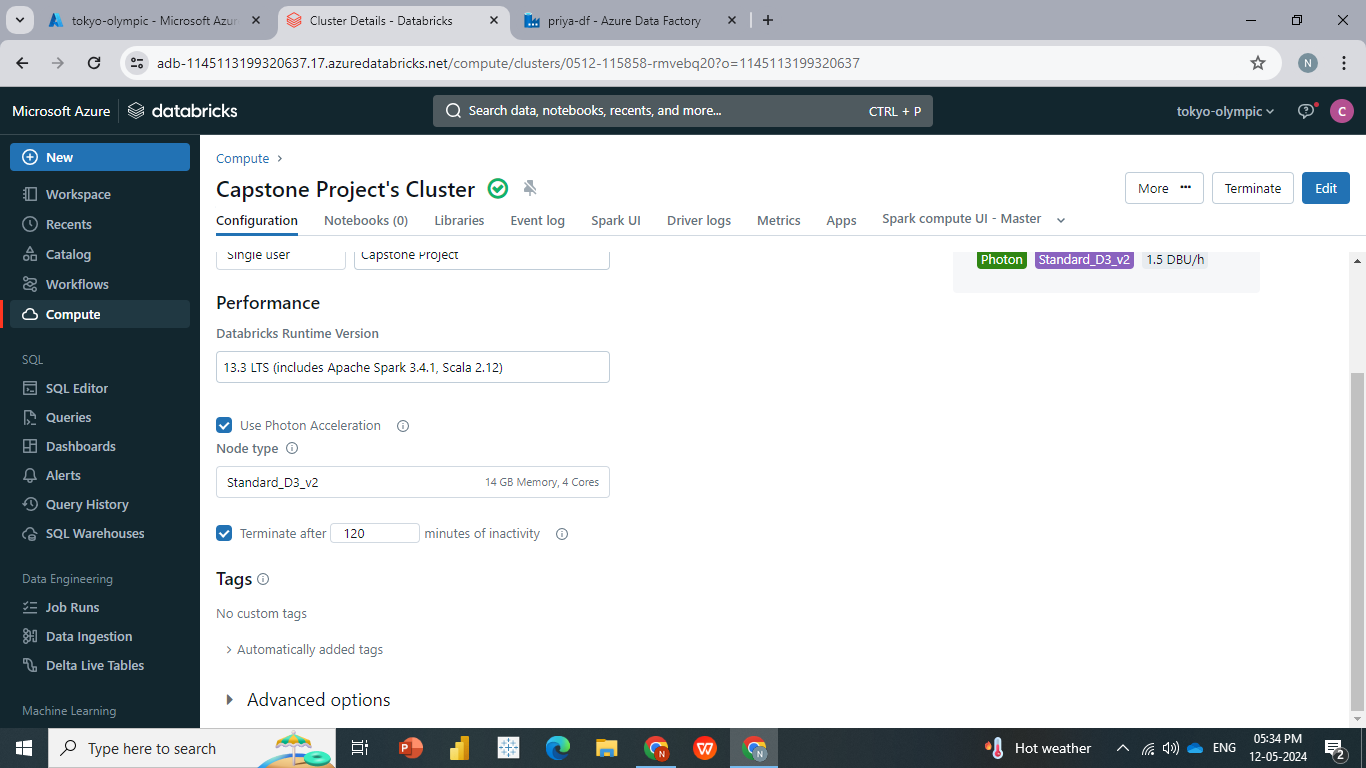
## **Data Transformation Process**

## **After the successful ingestion of data into the Data Lake, we deploy Azure Databricks notebooks. These notebooks attach to our storage account and perform the necessary data transformations. This process involves creating an app under the App Registration Service and providing it with access control to the storage account.**

* **Next, Go to Azure Databricks services and create an azure Databricks service, and launch studio and u will be redirected and sign-in into DataBricks platform.**



* **Create a computer instance, once our compute is ready, we can start coding and apply transformation as required.**
* **Once our Cluster is ready, we can Open new notebook and name it as Tokyo Olympic Transformation**
* **Search for App Registration, there are some steps involved we need to create an App and get some credentials to connect Azure Databricks to ADLS, and create a App as APP01 and register.**

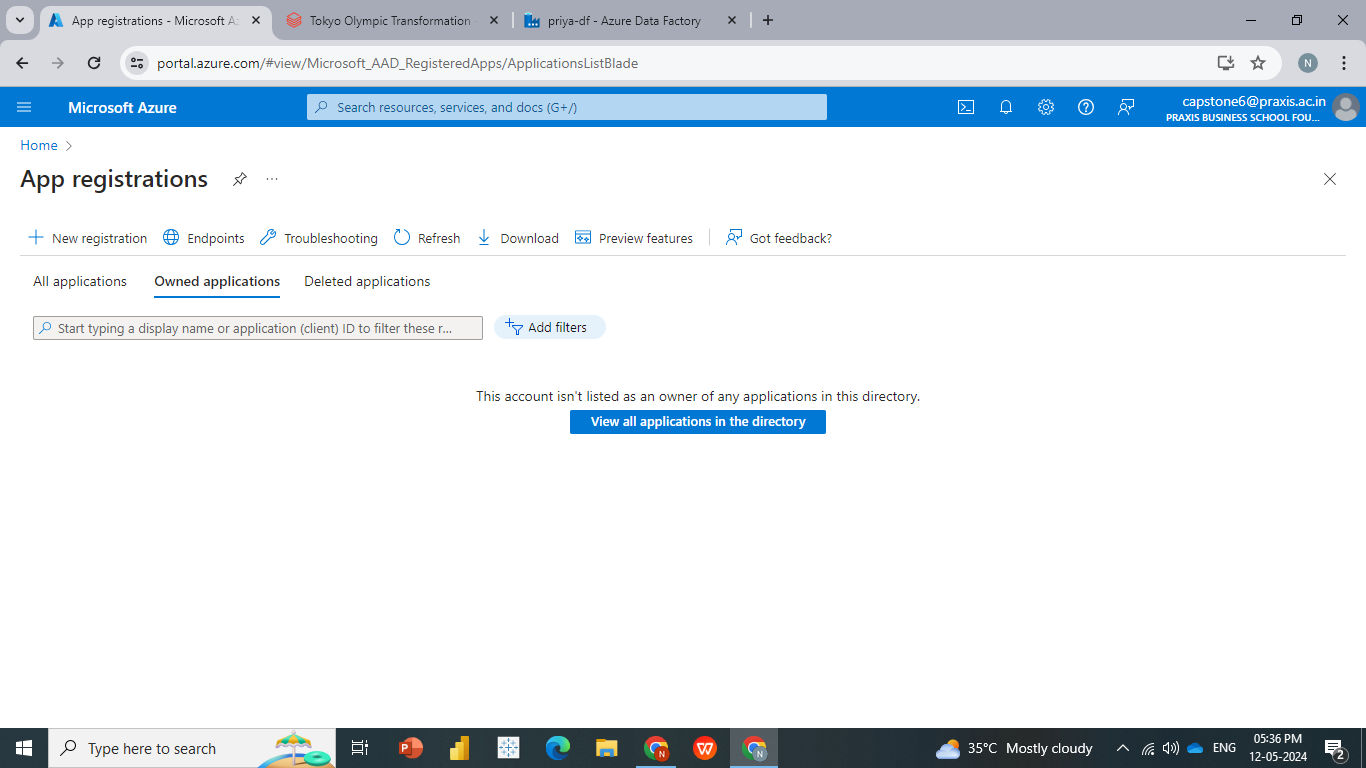


**Creating the app:**

* **Copy Client ID and Directory (Tenant Id) which we require for our Apache Spark and connect Azure Databricks to ADLS.**
* **Click on Certificates and secrets as we require to create a secret Id for accessing ADLS, click on new client secret --- > Name it as secret key --- > Add.**

[**Granting app access to the storage account:**](https://github.com/SherlockkOms/Tokyo-Olympics-Data-Analysis-Microsoft-Azure-End-to-End-Pipeline-Project?tab=readme-ov-file#granting-app-access-to-the-storage-account)

* **After this u can see that u have access to raw data as well as transformed data and u will be able to access all this data from different repositories.**
* **Paste all the credentials (Client ID and Directory (Tenant Id), Secret key) into the workspace.**
* **As shown below and test connection by Executing, this is the basic authentication we need to go through to create a connection from the Data Factory to Data Lake Storage.**



**Azure DataBricks Processing**

* **We explicitly need to give access to App01 to access data from Data Factory to Data Lake. For that we require to make some settings which are shown below.**

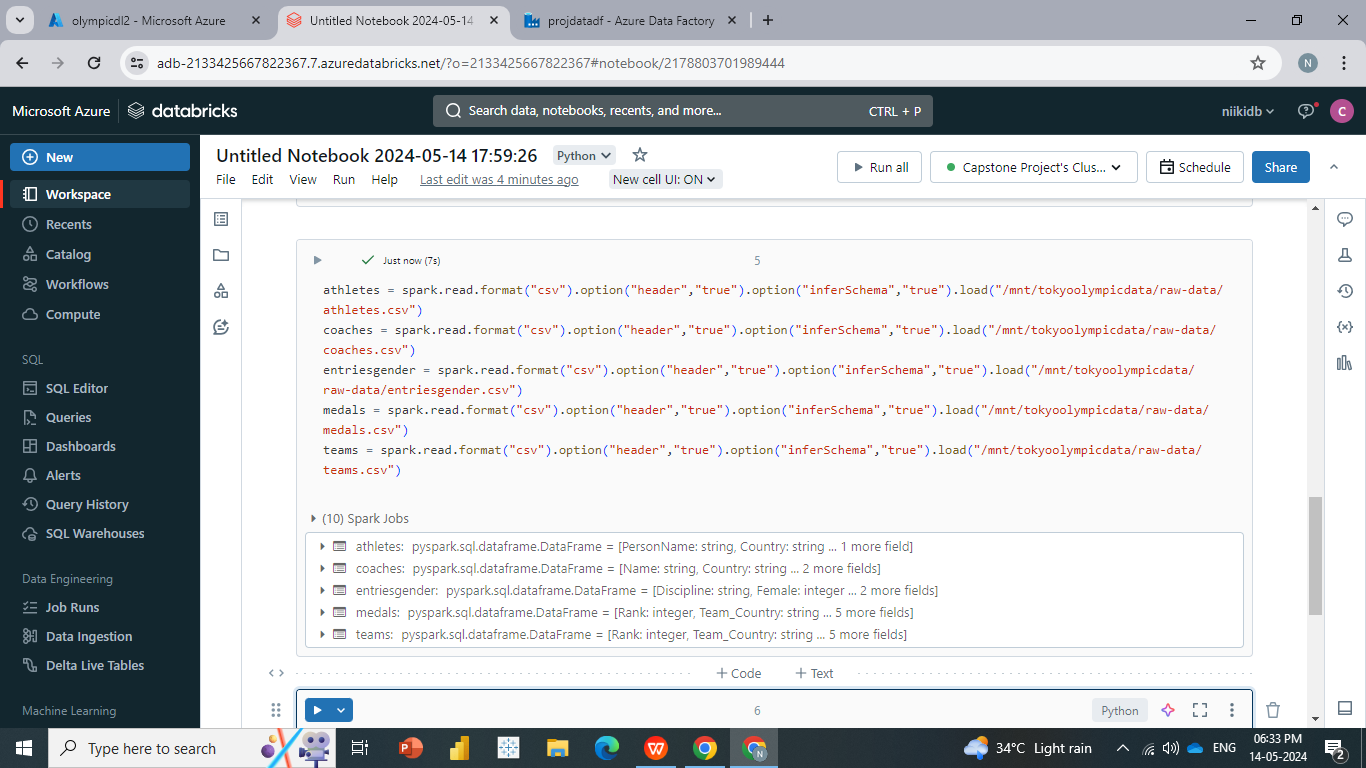
A screenshot of a computer

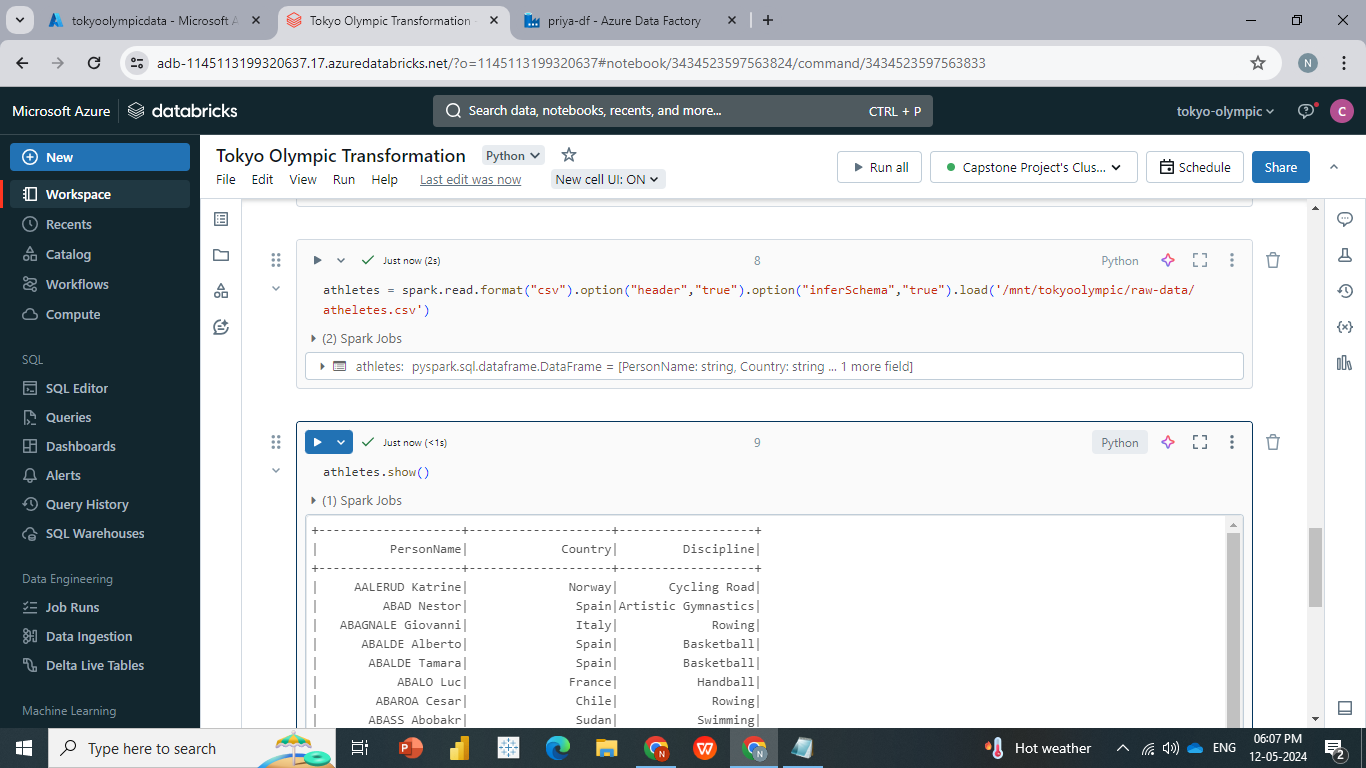
Description automatically generated

* **To read a file in Spark, we have syntax as:**

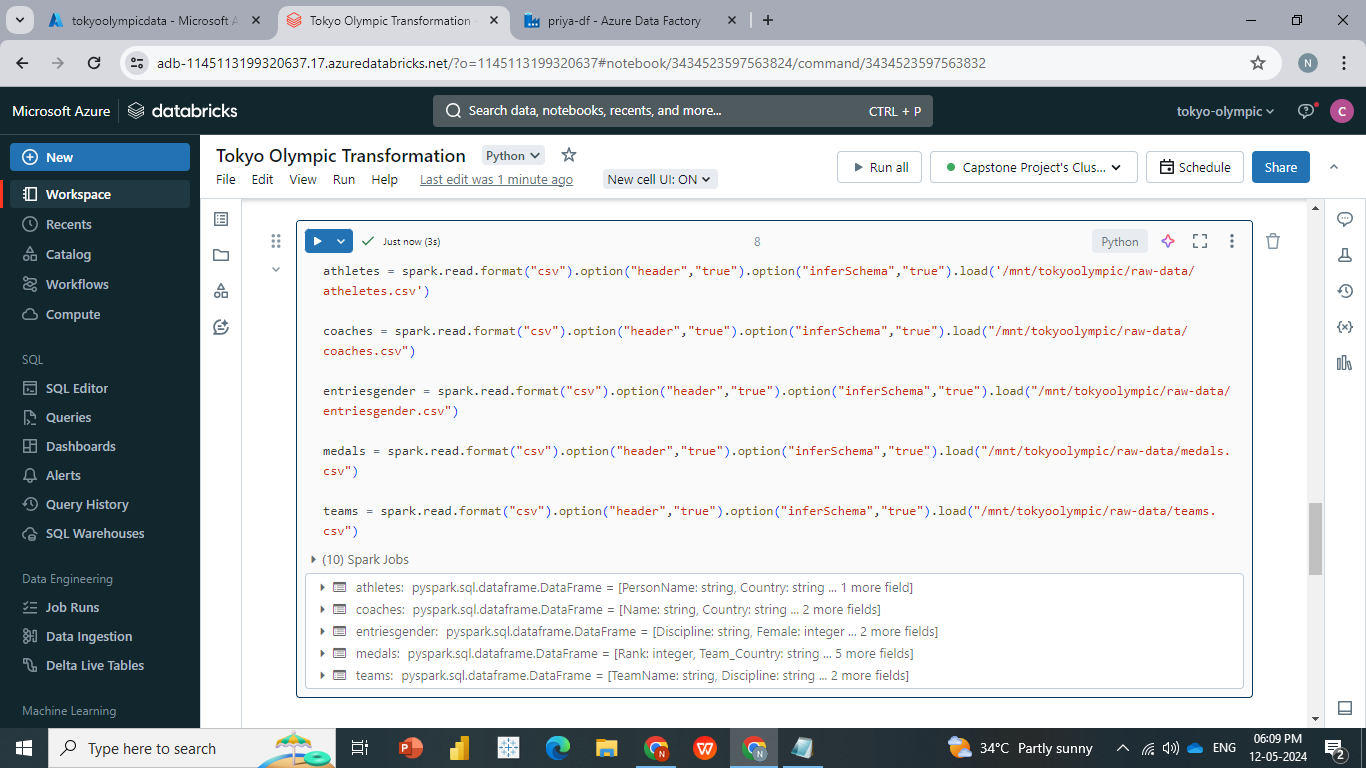
**df = spark.read.format("csv").option("header","true").load(filePath)**

**and to display output use df.show()**

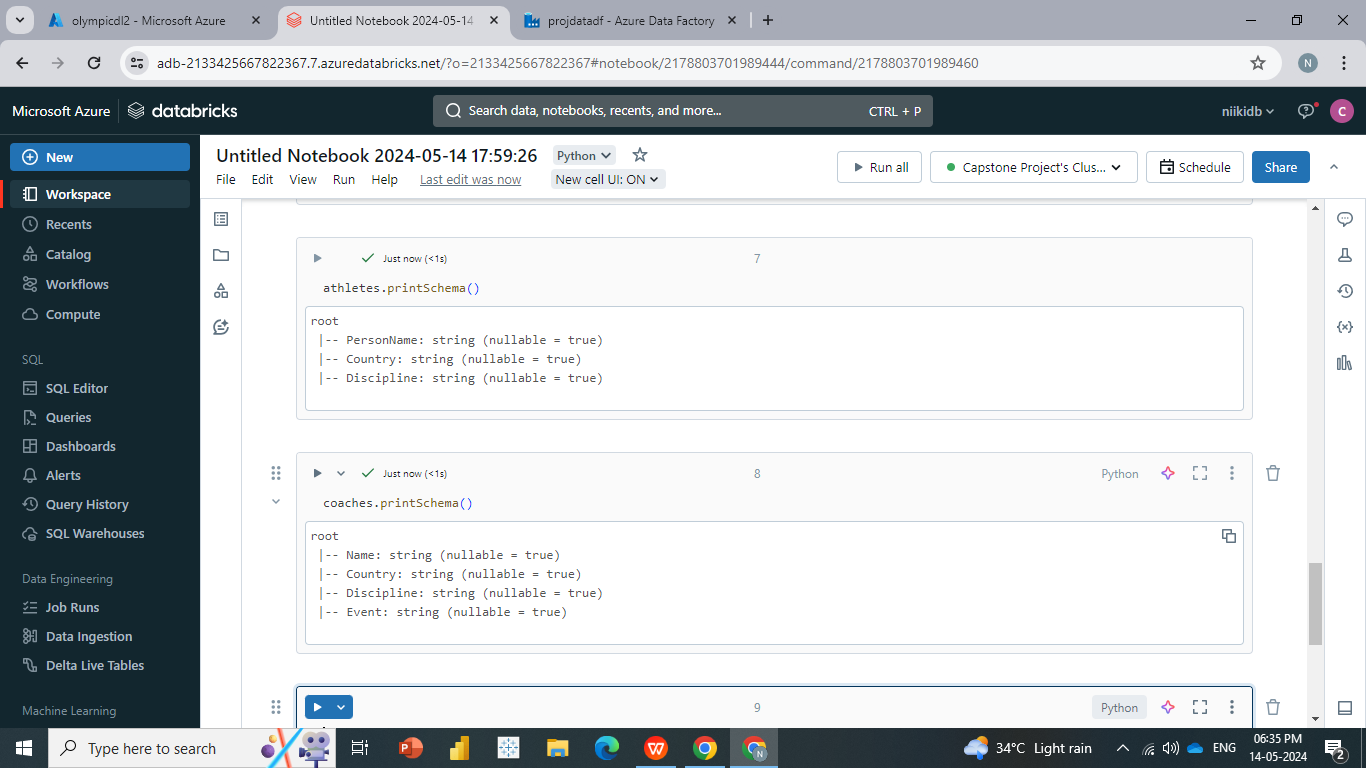




* **Read the date for the tables in using the same code.**

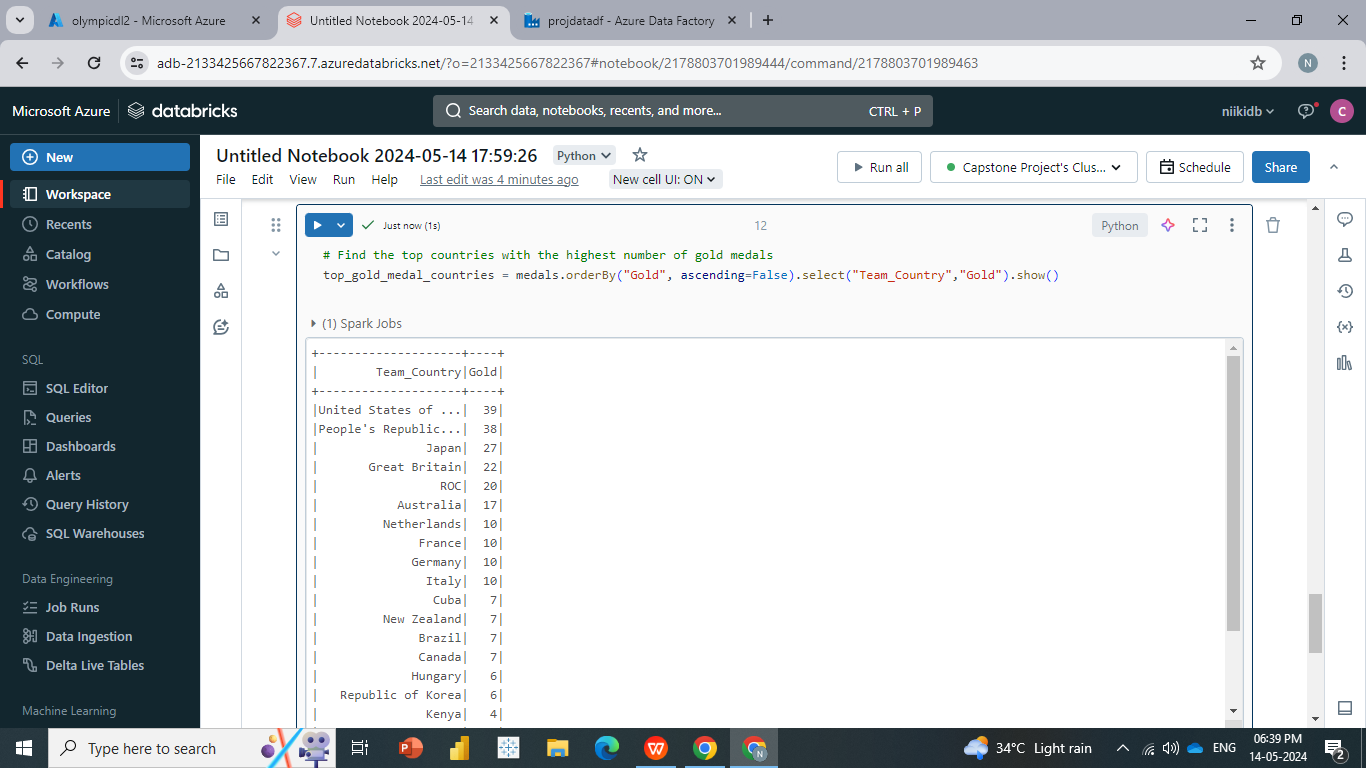


* **printSchema : printSchema() is used to print or display the schema of the DataFrame in the tree format along with column name and data type**
* **withColumn():  is a transformation function of DataFrame which is used to change the value, convert the datatype of an existing column,**

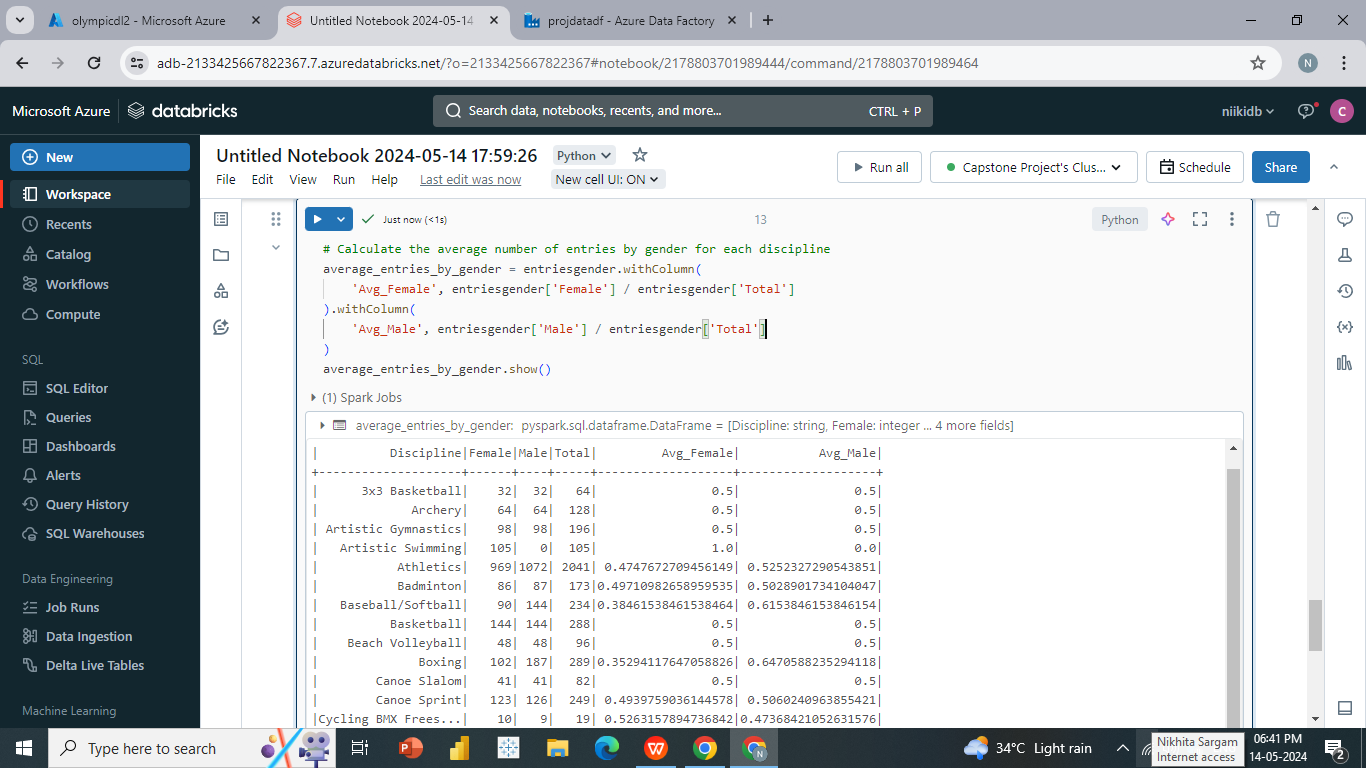


* **Carry out the basic transformation to validate the data and carry out some calculation.**

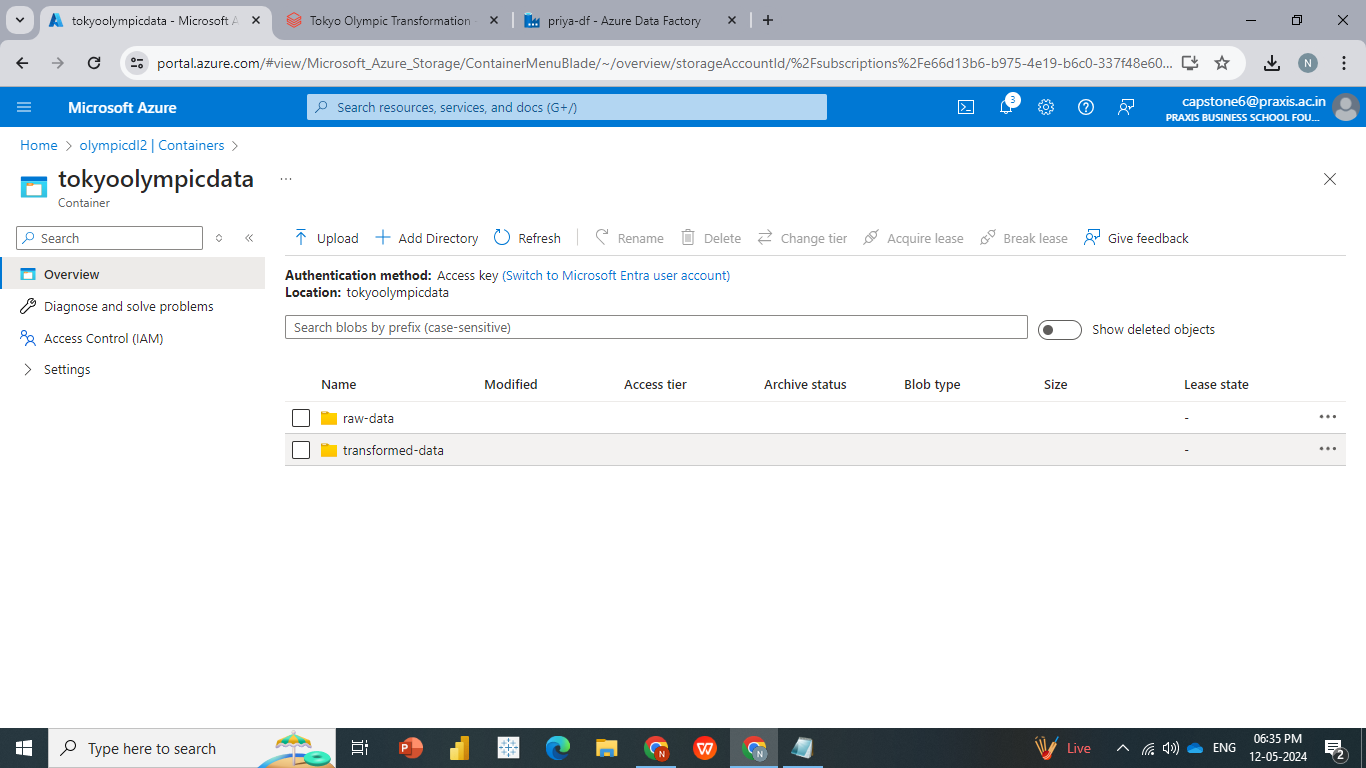
**Find the Top Countries with highest number of golds medals.**



**Calculate the average number of entries by gender for each discipline.**



**Then load back transferred data set to azure data factory.**



**EDA Analysis**

**The analysis involved loading two datasets from a data lake storage system. These datasets were then merged or joined together to create a unified dataset. Subsequently, an Exploratory Data Analysis (EDA) was conducted on the combined dataset spanning the years 1896 to 2016.**

**This analysis likely involved examining various aspects of the data such as trends, patterns, distributions, correlations, and outliers to gain insights into the historical Olympic data across different years.**

A screenshot of a computer

Description automatically generated

**Calculate the sum of null values for each column and display the result.**

A screenshot of a computer

Description automatically generated

**India details**

A screenshot of a computer

Description automatically generated**Japan details**

A screenshot of a computer

Description automatically generated

**Top 10 Countries**

**countries = [row['Team'] for row in top\_10\_countries]**

**counts = [row['count'] for row in top\_10\_countries]**

**plt.figure(figsize=(12, 6))**

**plt.title("Overall Participation by Country")**

**sns.barplot(x=countries, y=counts, palette="Set2")**

**plt.xticks(rotation=20)**

**plt.xlabel("Country")**

**plt.ylabel("Number of Participants")**

**plt.show()**

A graph of different colored bars

Description automatically generated

**Histogram of age distribution**

**A graph of age distribution

Description automatically generated**

**Plot the pie chart gender distribution.**

A pie chart with numbers and a number of people

Description automatically generated

**Total number of female athletes in each Olympic against each year**

A graph showing different colored bars

Description automatically generated

**Display the gold medal athlete details.**

A screenshot of a computer

Description automatically generated

**Plot the count of gold medals for athletes over 60 in the specified sporting event.**

**A graph of different colors and sizes

Description automatically generated**

**Get the top 5 countries with the most gold medals.**

A graph of different colors

Description automatically generated

**Count the number of gold medals won by each country.**

**A graph showing the top ten countries/regions with gold medals

Description automatically generated**

**Load the transformed data into target location with the help of below utility.**

A close-up of a text

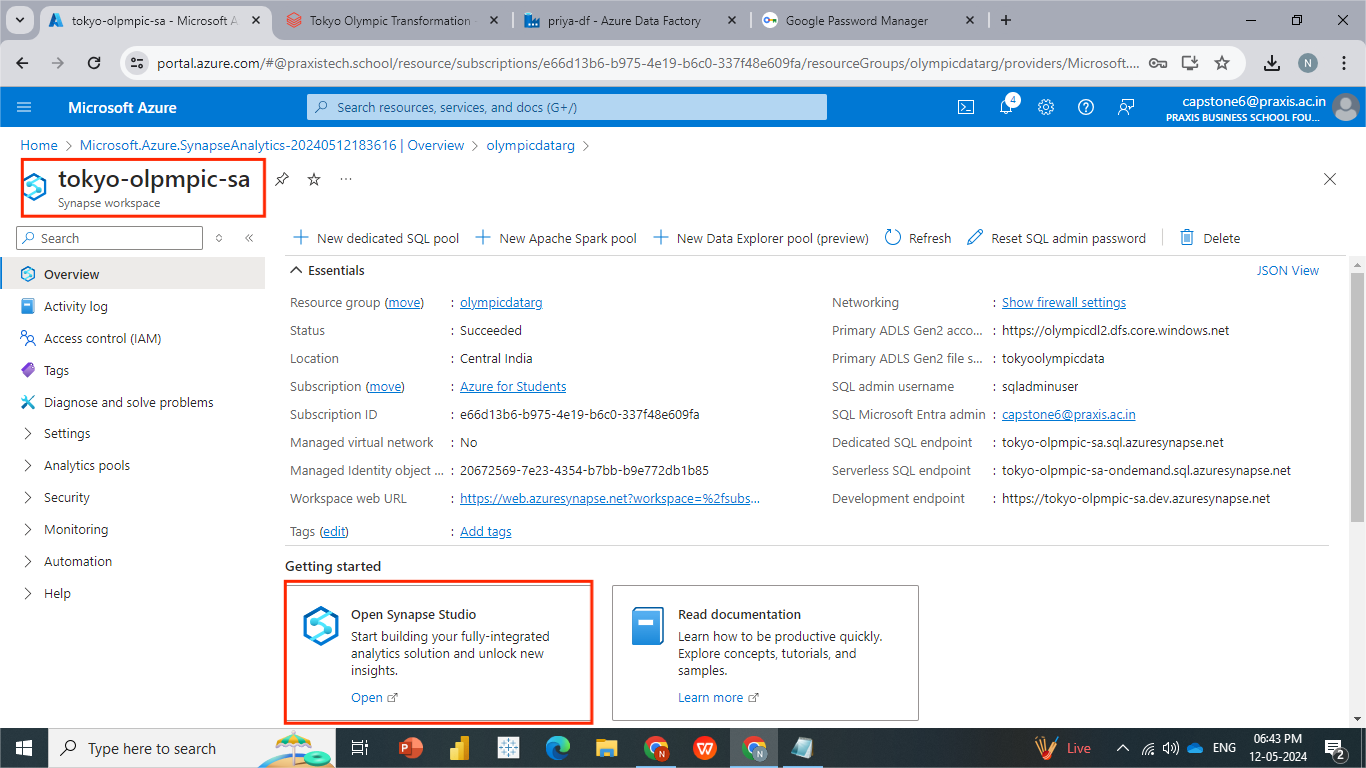
Description automatically generated

**Creation of the Database using Azure Synapse Analytics**

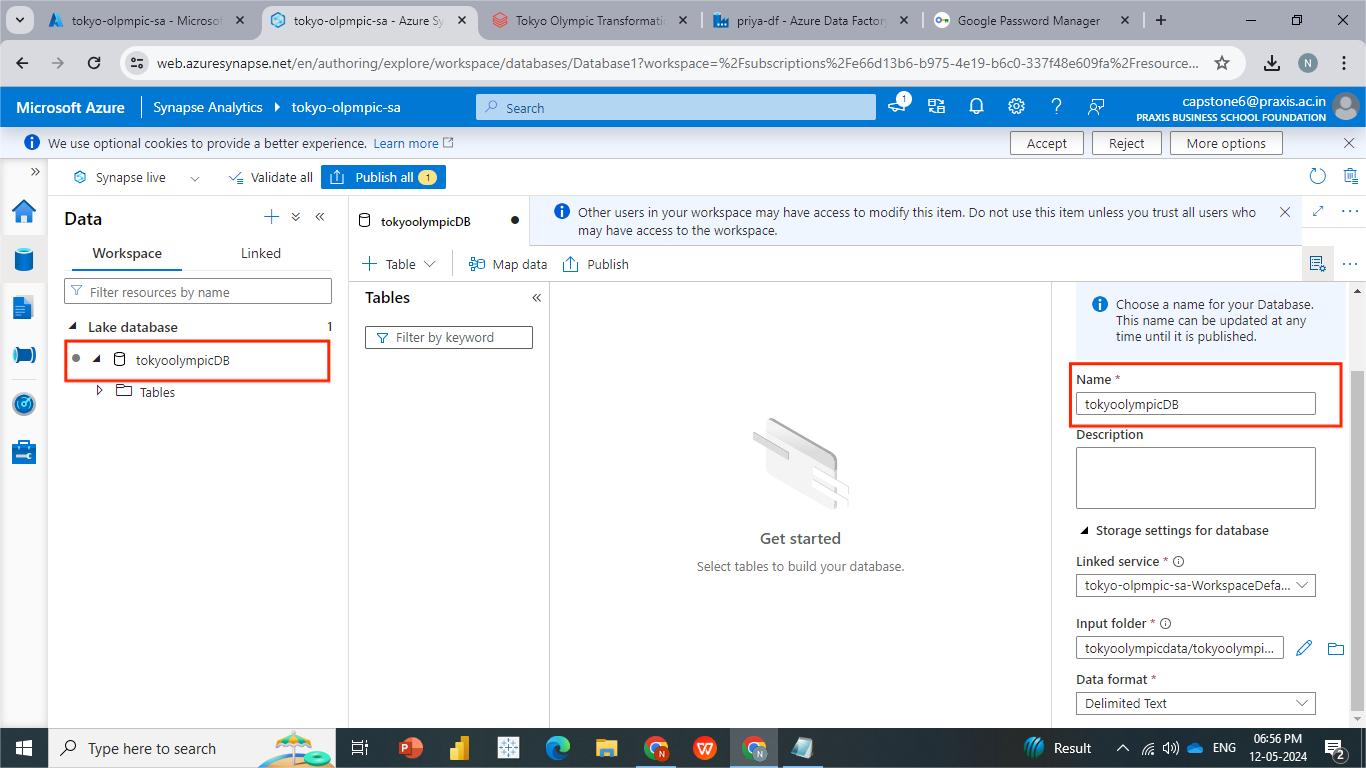
* **To create a database from our transformed data, we utilize Azure Synapse Analytics.**
* **Open Azure Synapse Analytics and create synapse workspace .**



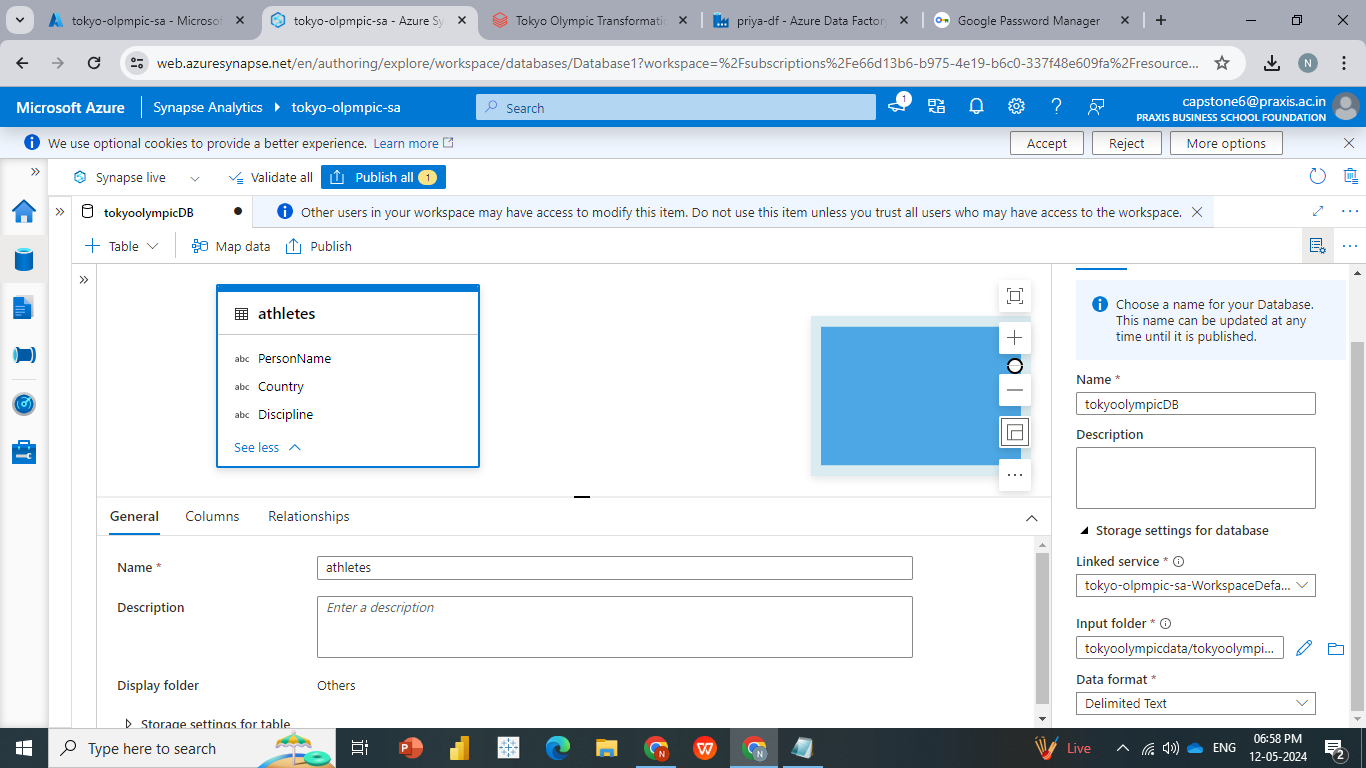
* **We can check under olympicdatarg all our services are stored under resource group as shown**
* **Open Synapse studio, it will be redirected automatically once you click on the open studio**



* **Open Data --- > Choose Lake database ---- > Name it as per convenience.**



* **Once our database is created, we can load tables into our database.**
* **Now click on tables ---> from Data lake ---> Provide External Name (Athlete) ----> choose default linked services --- > choose folder / file location (Tokyo-Olympic-data / transformed-data / athletes/choose single file) ----> create ----> validate ----> publish**

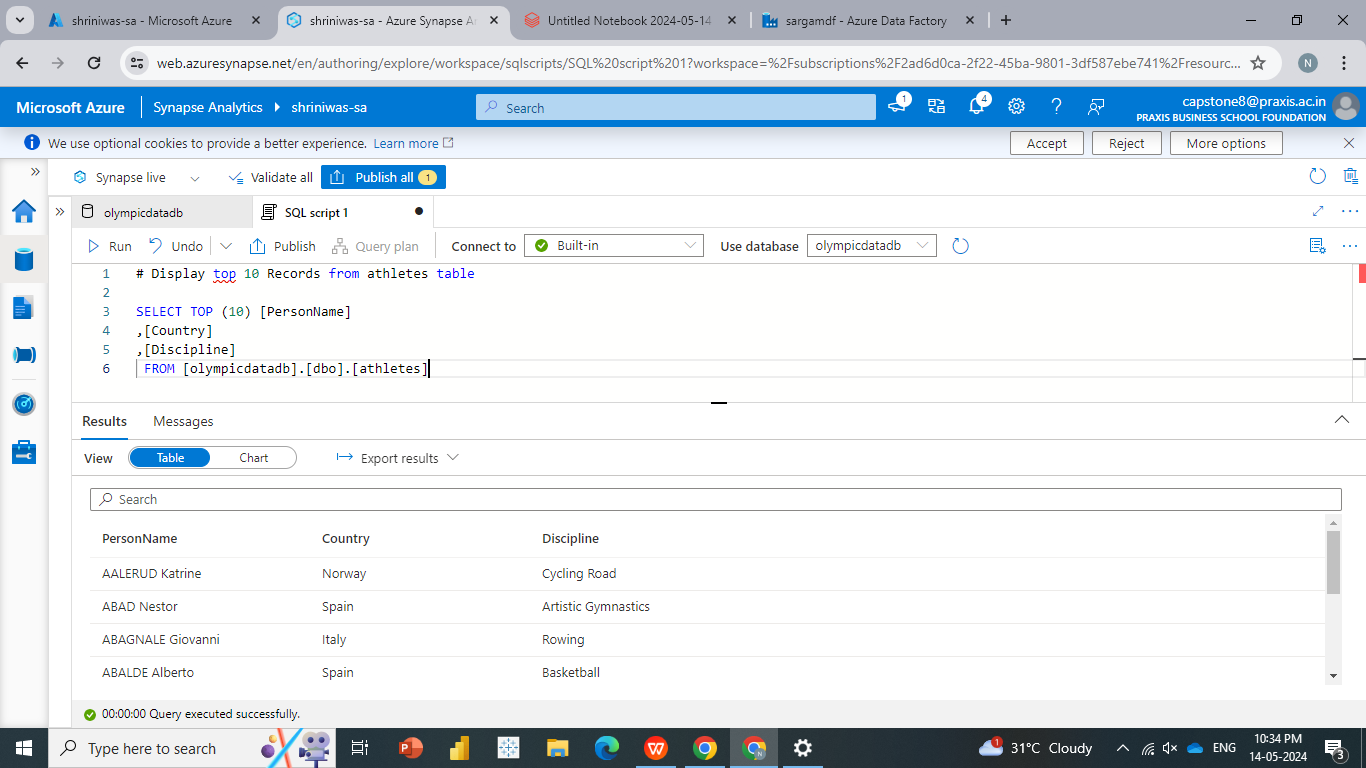


* **Similarly add all other transferred data into it**

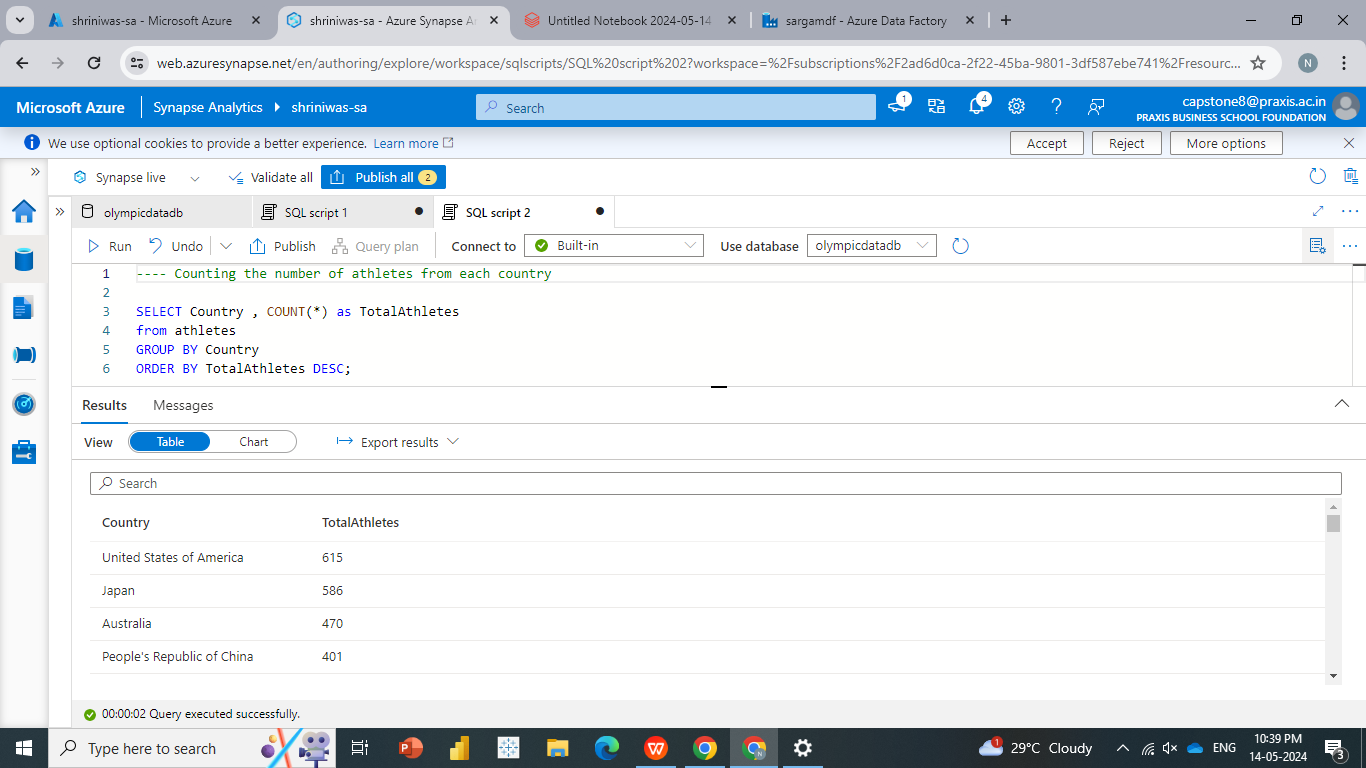


**Now we will perform some Analytics using SQL Queries in Azure Synapse Analytics**

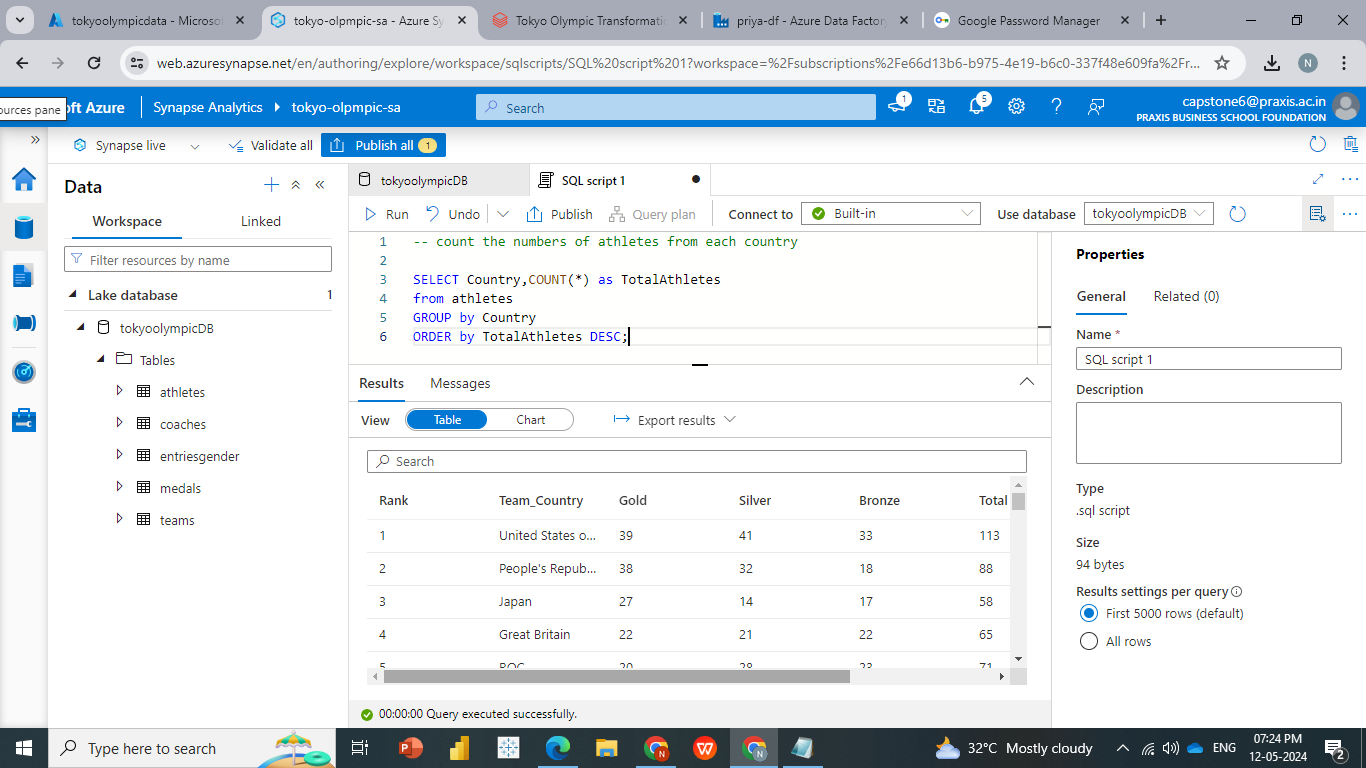
**Display top 10 records from Athletes table.**



* 1. **Counting the number of athletes from each country.**

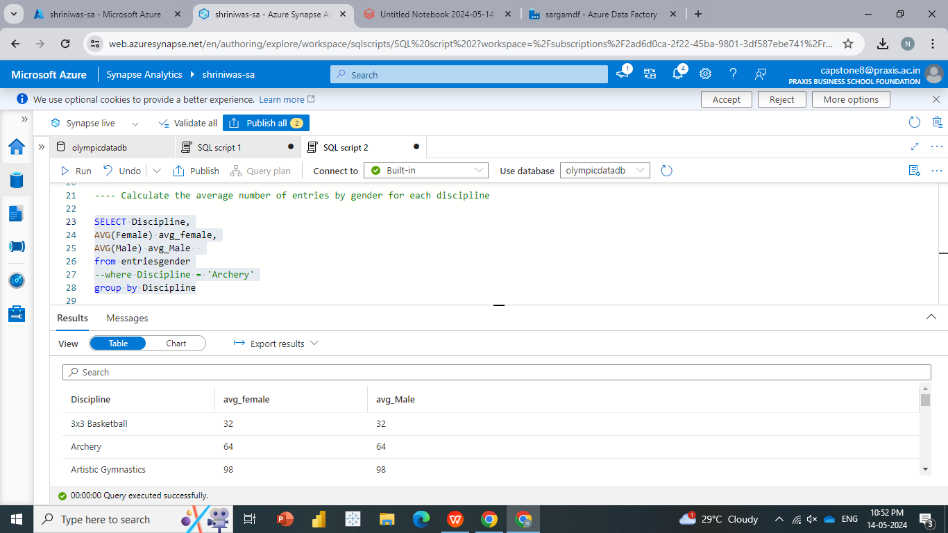


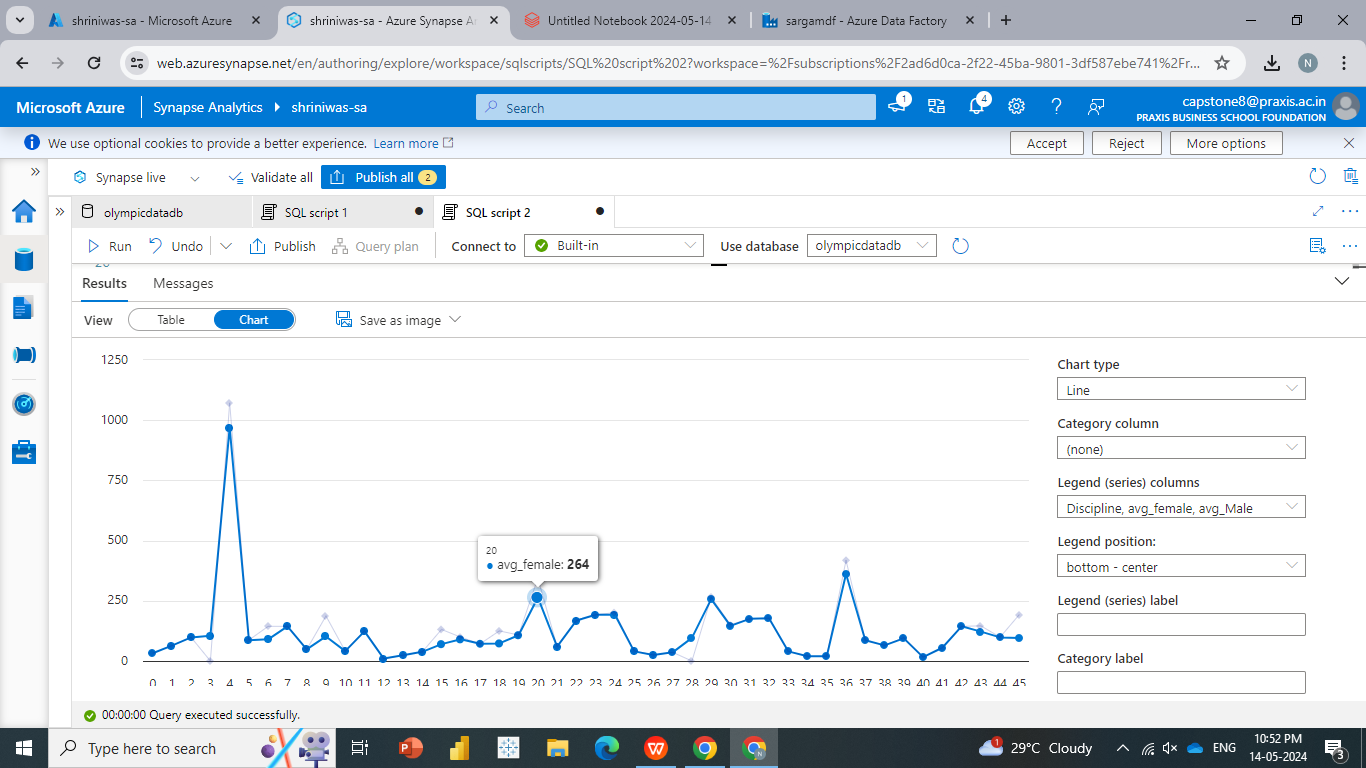
**Count number of athletes from each country**





**Calculate the average number of entries by gender for each discipline.**



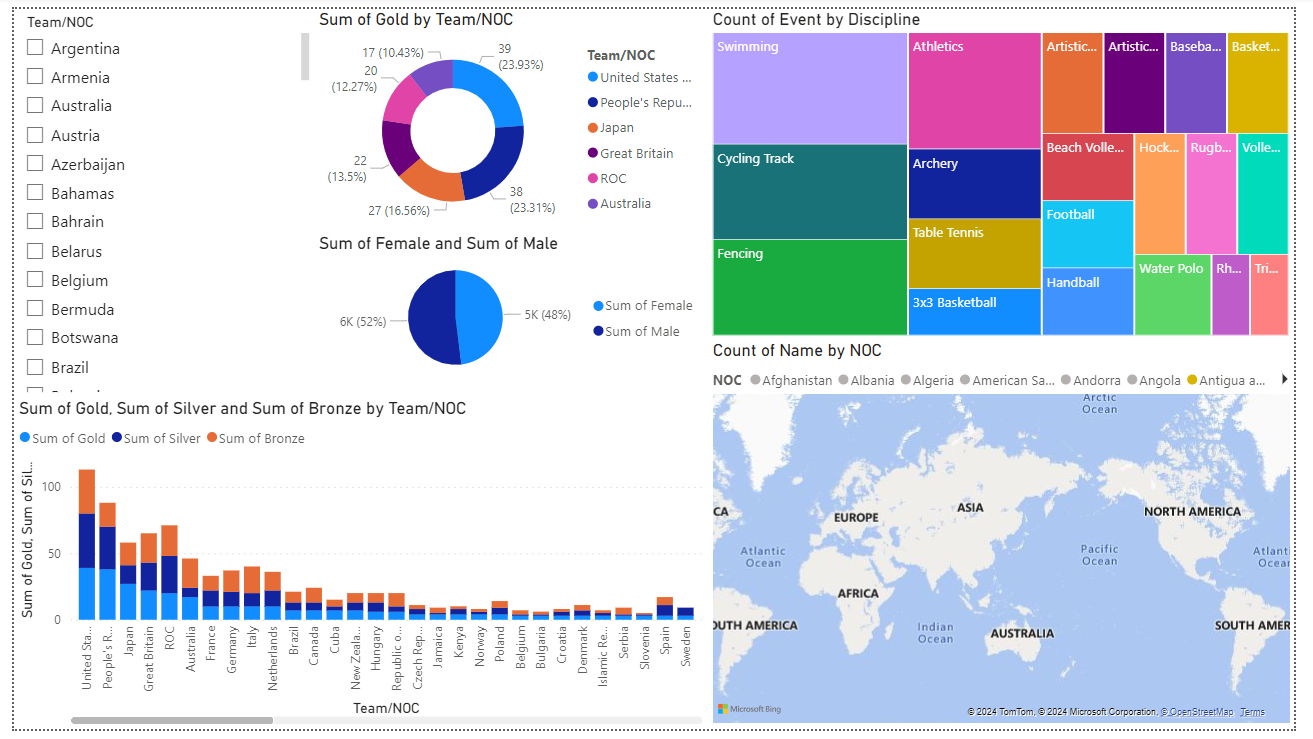


[**Generating a report in Power-BI**](https://github.com/SherlockkOms/Tokyo-Olympics-Data-Analysis-Microsoft-Azure-End-to-End-Pipeline-Project?tab=readme-ov-file#generating-a-report-in-powerbi)

* **Load the data into Power BI and transfer the data into required to plot the visualization.**
* **Below are the visualizations obtained from the Olympic data 2019.**
* And .pbix file has been attached for the reference.



**Visualization View**



**Future Scope and Conclusion**

**Future Scope**

* **Advanced Analytics: Incorporate advanced analytics techniques such as machine learning and predictive modelling to forecast future trends in Olympic data. This could involve predicting medal counts, athlete performance, or audience engagement metrics.**

* **Real-time Data Processing: Explore real-time data processing capabilities to enable near-instantaneous analysis and decision-making during ongoing Olympic events. This could involve streaming data ingestion and processing using services like Azure Stream Analytics.**
* **Integration with External Data Sources: Extend the pipeline to incorporate data from external sources such as social media, weather forecasts, or economic indicators to provide a comprehensive analysis of factors influencing the Olympics.**
* **Enhanced Visualization: Enhance the visualization aspect by incorporating interactive dashboards, geospatial analysis, and storytelling features to make the insights more engaging and accessible to a wider audience.**
* **Data Governance and Compliance: Implement robust data governance and compliance measures to ensure data security, privacy, and regulatory compliance, especially considering the sensitive nature of Olympic data.**

**Conclusion**

**The development of an end-to-end data pipeline for managing historic Olympic data represents a significant milestone in showcasing the power of cloud computing and Azure's data services. Through this project, valuable insights have been extracted, highlighting the significance of analysing historical Olympic data. The integration of Azure Databricks, Azure Data Factory, and other Azure resources has ensured seamless operation and scalability, enabling efficient processing and analysis of large volumes of data.**

**Moving forward, the project has the potential to evolve further, with opportunities to incorporate advanced analytics, real-time processing, integration with external data sources, and enhanced visualization techniques. By continually refining the pipeline and expanding its capabilities, we can unlock even greater insights into the world of the Olympics, contributing to our understanding of sports, culture, and society on a global scale.**