Data Engineering Project 1

Real-Time Analytics Dashboard Name: Pradip Bochare

(Tokyo Olympics)

1.1 Project Statement

Build a real-time analytics dashboard using Azure Synapse Analytics for data storage, Azure Databricks for real-time analytics.

1.2 Project Overview

Welcome to the Tokyo-Olympics Data Engineering project on Azure! This project is designed to showcase how various Azure services can be utilized to ingest, store, transform, and analyse Olympics-related data. This project provides a data engineering and analytical journey on the Tokyo Olympic dataset. Starting with a CSV on GitHub, the data is ingested into the Azure ecosystem via Azure Data Factory. It's initially stored in Azure Data Lake Storage Gen2, then transformed in Azure Databricks. The enriched data, once again housed in ADLS Gen2, undergoes advanced analytics in Azure Synapse. The insights are finally visualized in Azure Synapse or Power BI, offering a comprehensive view of the dataset.

1.3 Project Requirement:

The project aims to develop a real-time analytics dashboard leveraging various Azure services including ADLS Gen2, Azure Synapse Analytics, Azure Databricks, Azure Data Factory, and Power BI. The dashboard will enable users to visualize and analyse data in real-time, providing actionable insights for informed decision - making.

A. Data Storage with ADLS Gen2 and Azure Synapse Analytics

1) Utilize Azure Synapse Analytics as the central data storage solution**.**
2) Implement data lakes and SQL pools to accommodate structured and unstructured data.
3) Design storage structures optimized for efficient querying and analysis of real-time data stream
4) Utilize Azure Data Lake Storage Gen2 (ADLS Gen2) to store large volumes of data efficiently and securely.
5) Design storage structures optimized for efficient querying, analysis, and storage of real-time data streams

B. Real-Time Analytics with Azure databricks

1) Leverage Azure Databricks for real-time analytics processing of streaming data
2) Develop Spark-based jobs or notebooks within Azure Databricks to analyze streaming data and used for transforming the data

C. **Data Orchestration with Azure Data Factory**

1) Implement Azure Data Factory to orchestrate data pipelines for ingesting, transforming, and loading data into Azure Synapse Analytics and Azure Databricks.
2) Utilize Data Factory pipelines to handle complex data transformations and data movement tasks efficiently.
3) Configure event-driven triggers to initiate data processing and analytics workflows in real-time.

D. Dashboard Development with PowerBI

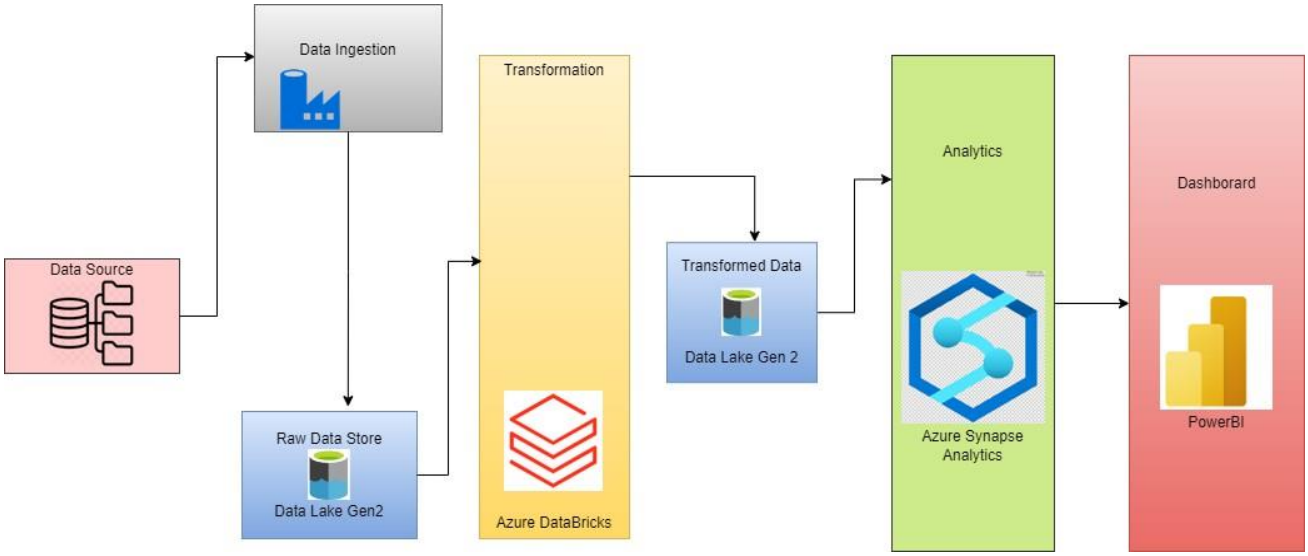
1) Design and develop an interactive dashboard interface using Power BI for visualizing real-time analytics insights.
2) Connect Power BI to Azure Synapse Analytics and Azure Databricks to fetch live data streams and analytics results.
3) Connect Power BI to Azure Synapse Analytics and Azure Databricks to fetch live data streams and analytics.
4) Ensure the dashboard is responsive, intuitive, and user-friendly, allowing users to explore and interact with data effectively.

E. Integration and data connectivity

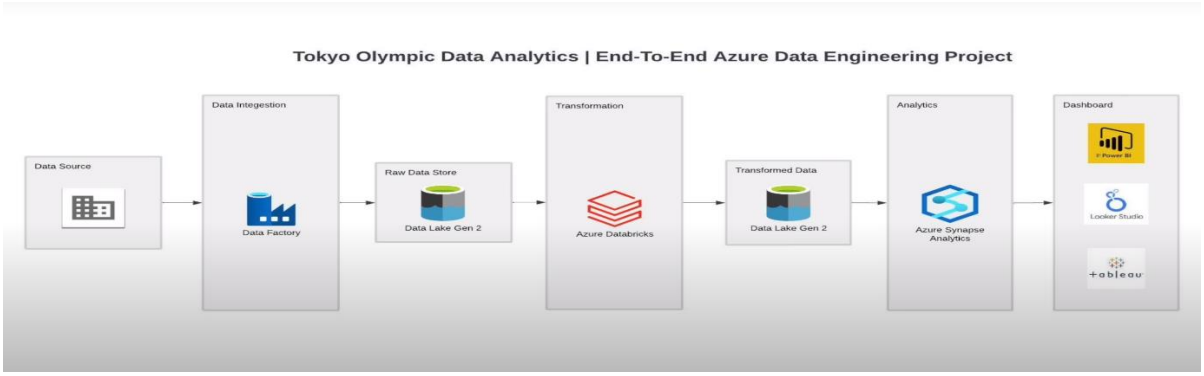
1) Integrate the real-time analytics dashboard with Azure Databricks and Azure Synapse Analytics to fetch live data streams and analytics results.
2) Establish secure connections between Power BI and backend data sources to ensure data integrity and confidentiality.
3) Implement efficient data refresh mechanisms to keep the dashboard up-to-date with the latest analytics insights.

2.1 Architectural Diagram

The architecture for the project involving building a real-time analytics dashboard using Azure Synapse Analytics for data storage, Azure Databricks for real-time analytics, Azure Data Factory for orchestrating data pipelines, and Power BI for dashboard visualization, along with ADLS Gen2.



End to End pipeline architecture in azure



3.1 Tools Used in Real time Analytics Dashboard

To build a real-time analytics dashboard using ADLS gen 2 Azure Synapse Analytics for data storage, Azure Databricks for real-time analytics, Azure Data Factory for orchestrating data pipelines, and Power BI for dashboard visualization, the following tools and services are utilized:

1)Azure Data Lake Storage Gen2(ADLS Gen2)

ADLS Gen2 provides scalable, secure, and cost-effective storage for big data analytics workloads. It allows storage of structured, semi-structured, and unstructured data in a distributed file system. ADLS Gen2 integrates seamlessly with other Azure services, making it suitable for real-time analytics scenarios.

2)Azure Synapse Analytics

Formerly known as Azure SQL Data Warehouse, Azure Synapse Analytics is a cloud-based analytics service that brings together enterprise data warehousing and big data analytics. It provides capabilities for data storage, data integration, data warehousing, and big data analytics in a single platform. Azure Synapse Analytics enables the storage and analysis of large volumes of data in real-time, making it ideal for building real-time analytics dashboards.

3)Azure Databricks

Azure Databricks is an Apache Spark-based analytics platform optimized for big data and machine learning workload. It provides a collaborative environment for data engineers, data scientists, and analysts to work together on data analytics projects. Azure Databricks enables real-time analytics processing of streaming data and supports various analytics and machine learning tasks.

4)Azure Data Factory

Azure Data Factory is a cloud-based data integration service that allows users to create, schedule, and orchestrate data pipelines. It supports the movement and transformation of data between various sources and destinations, including ADLS Gen2, Azure Synapse Analytics, and Azure Databricks. Azure Data Factory enables the creation of real-time data pipelines for ingesting, processing, and transforming data for analytics purposes.

5)PowerBI

Power BI is a business analytics service provided by Microsoft for creating interactive dashboards and reports. It connects to various data sources, including Azure Synapse Analytics, ADLS Gen2, and other Azure services, to fetch real-time data streams and analytics results. Power BI offers user-friendly interface for designing and customizing dashboards with rich visualizations, charts, and graphs.

Dataset Used

This contains the details of over 11,000 athletes, with 47 disciplines, along with 743 Teams taking part in the 2021(2020) Tokyo Olympics. This dataset contains the details of the Athletes, Coaches, Teams participating as well as the Entries by gender. It contains their names, countries represented, discipline, gender of competitors, name of the coaches.

Prerequisite

1. **Azure Data Factory:** For data ingestion from GitHub.
2. **Azure Data Lake Storage Gen2**: As the primary data storage solution.
3. **Azure Databricks:** For data transformation tasks.
4. **Azure Synapse Analytics:** To perform in-depth data analytics.

3.2 Execution Overview

The execution process for building a real-time analytics dashboard using ADLS Gen2, Azure Synapse Analytics, Azure Databricks, Azure Data Factory, and Power BI:

1)Azure Environment Setup

Provision Azure services including ADLS Gen2, Azure Synapse Analytics, Azure Databricks, Azure Data Factory, and Power BI. Configure networking, security, and access controls based on project needs.

2)Data Ingestion with Data Factory

Create data ingestion pipelines in Azure Data Factory to ingest data from various sources into ADLS Gen2 and Azure Synapse Analytics. Implement real-time data ingestion where applicable.

3)Transform data with Azure Databricks

Develop real-time analytics solutions using Azure Databricks to process data streams and derive insights. Implement data transformations, aggregations, and machine learning algorithms for analytics.

4)Data store and management

Optimize data storage structures and configurations in ADLS Gen2 and Azure Synapse Analytics. Implement partitioning, indexing, and compression techniques for efficient storage and querying.

5)Dashboard development with PowerBI

Connect Power BI to Azure Synapse Analytics, ADLS Gen2, and other data sources. Design interactive dashboards with Power BI's visualization tools. Create reports, charts, graphs, and other visualizations to display real-time analytics insights.

4.1 Implementation and Development Process

To get started with this project, follow these steps:

1) Infrastructure setup and configuration: Provision Azure services including ADLS Gen2, Azure Synapse Analytics, Azure Databricks, Azure Data Factory, and PowerBI. Configure networking, security, and access controls based on project requirements.

2) **Data Modelling and Schema Design:** Design data models and schemas for storing and processing data in ADLS Gen2 and Azure Synapse analytics. Define data structures and formats for efficient real-time analytics processing

3) **Data Ingestion**: Use Azure Data Factory to configure data ingestion from your chosen data sources. Define the data movement and transformation activities required to bring the Olympics data into the Azure ecosystem. Implement data ingestion pipelines using Azure Data Factory to ingest data from multiple sources including IoT devices, applications, and databases into ADLS Gen2 and Azure Synapse. Ensure real-time or near real-time data ingestion capabilities to enable timely analytics

4) **Data Transformation**: Utilize Azure Databricks with PySpark for data transformation. Cleanse, preprocess, and reshape the data as necessary to prepare it for analysis. Develop Spark-based jobs or notebooks to analyze streaming data, perform transformations, and derive insights in real-time.

5) **Data Storage and management optimization**: Optimize data storage structures and configurations in ADLS Gen2 and Azure Synapse Analytics for efficient querying, storage, and retrieval. Implement partitioning, indexing, and compression techniques to improve performance and scalability.

6) **Analytics**: Write SQL queries using Azure Synapse Analytics to gain insights from the transformed data. Identify trends, statistics, and patterns related to the Tokyo-Olympics data.

7) **Dashboard Development with PowerBI**: Connect Power BI to Azure Synapse Analytics, ADLS Gen2, and other data sources to retrieve real-time data streams and analytics results. Design interactive dashboards with Power BI's visualization tools to present insights derived from real-time. Customize dashboard layouts, visualizations, and interactivity features based on user requirement. Connect to PowerBI for visualization Process.

8) **Integration and connectivity**: Establish seamless integration between Azure Databricks, Azure Synapse Analytics, ADLS Gen2, Azure Data Factory, and PowerBI. Configure data connectors, APIs, and authentication mechanisms for secure data exchange and connectivity.

9) **Testing and Quality Assurance**: Conduct comprehensive testing of data pipelines, analytics algorithms, and dashboard functionality. Perform unit tests, integration tests, and end-to-end tests to validate data accuracy, system performance, and user experience.

10) **Development and production rollout**: Deploy the real-time analytics dashboard to the production environment. Monitor system performance, data integrity, and user interactions during the initial rollout phase. Address any issues or bugs identified during deployment and ensure smooth operation of the dashboard.

11) **Documentation and User Training**: Document the architecture, design decisions, configuration settings, and implementation details of the real-time analytics dashboard. Provide user documentation and training materials to educate users on how to interact with the dashboard and interpret analytics insights effectively.

By following this implementation and development process, you can build a robust real-time analytics dashboard that leverages the capabilities of ADLS Gen2, Azure Synapse Analytics, Azure Databricks, Azure Data Factory, and Power BI to enable data-driven decision-making and insights generation in real-time

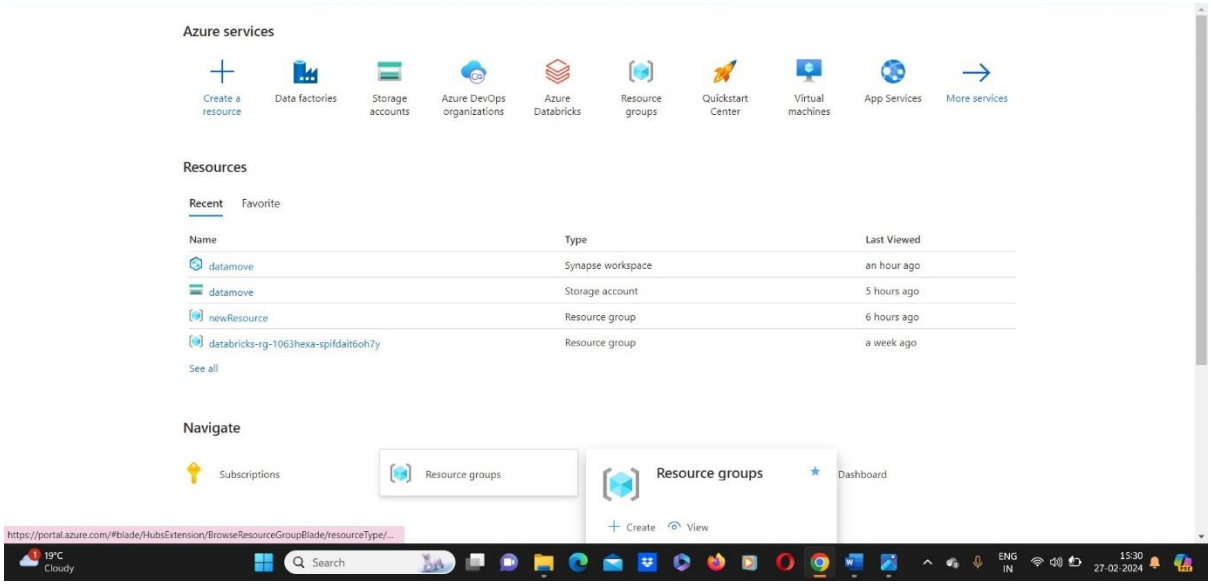
5.1 Tasks Performed on Real-time analytics Project

I used the dataset from Olympics 2021 In Tokyo and tried to use every important feature Azure had to offer. I **integrated** the data from my computer to a **data lake** by using **Microsoft Azure Data Factory**. A pipeline was built to integrate all data and **validated** to check for any errors.

Workflow

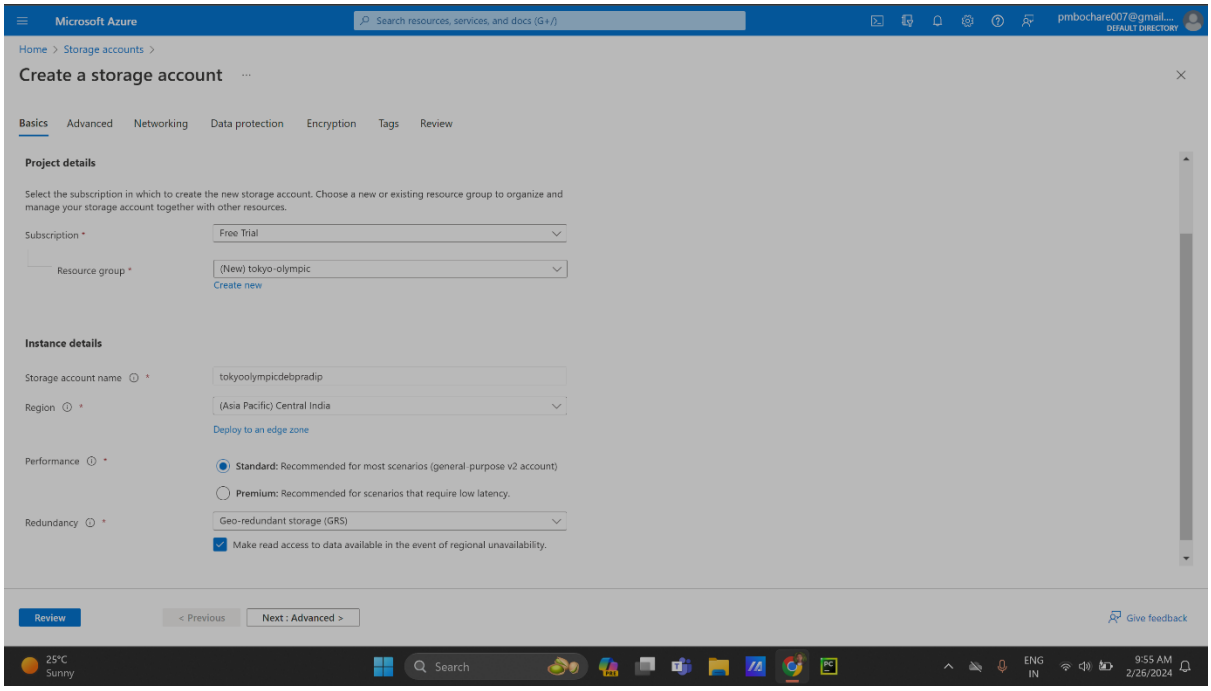
Initial Setup

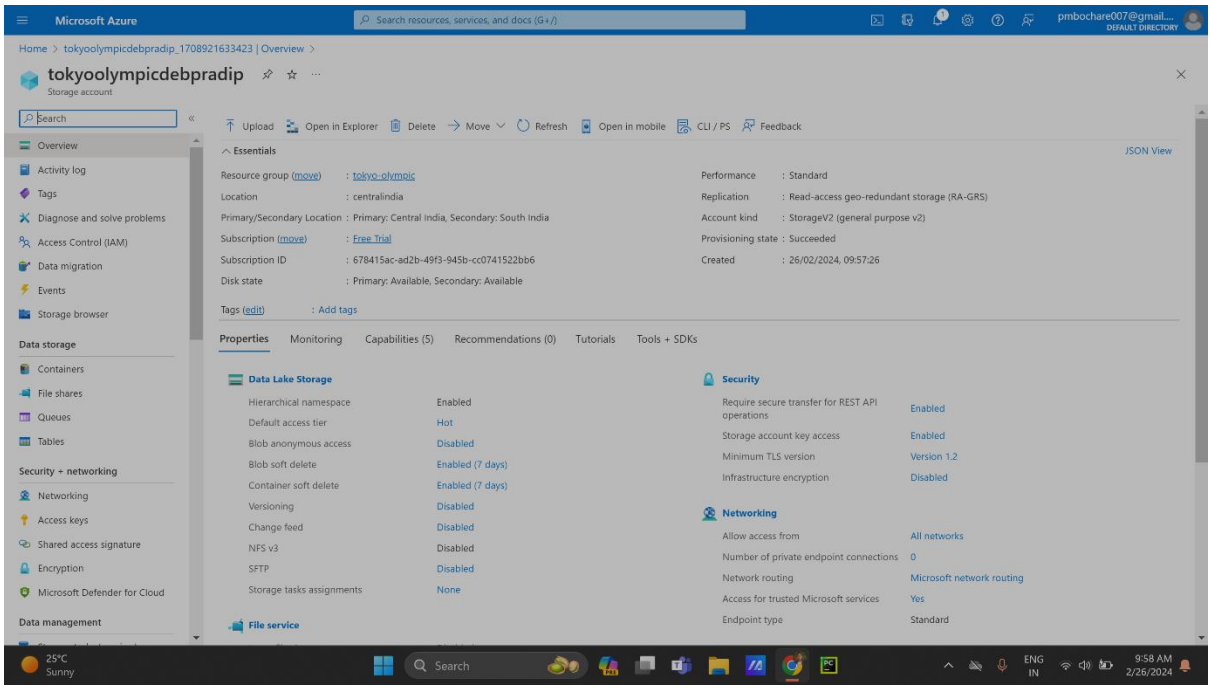
1. Create Azure Free Subscription account



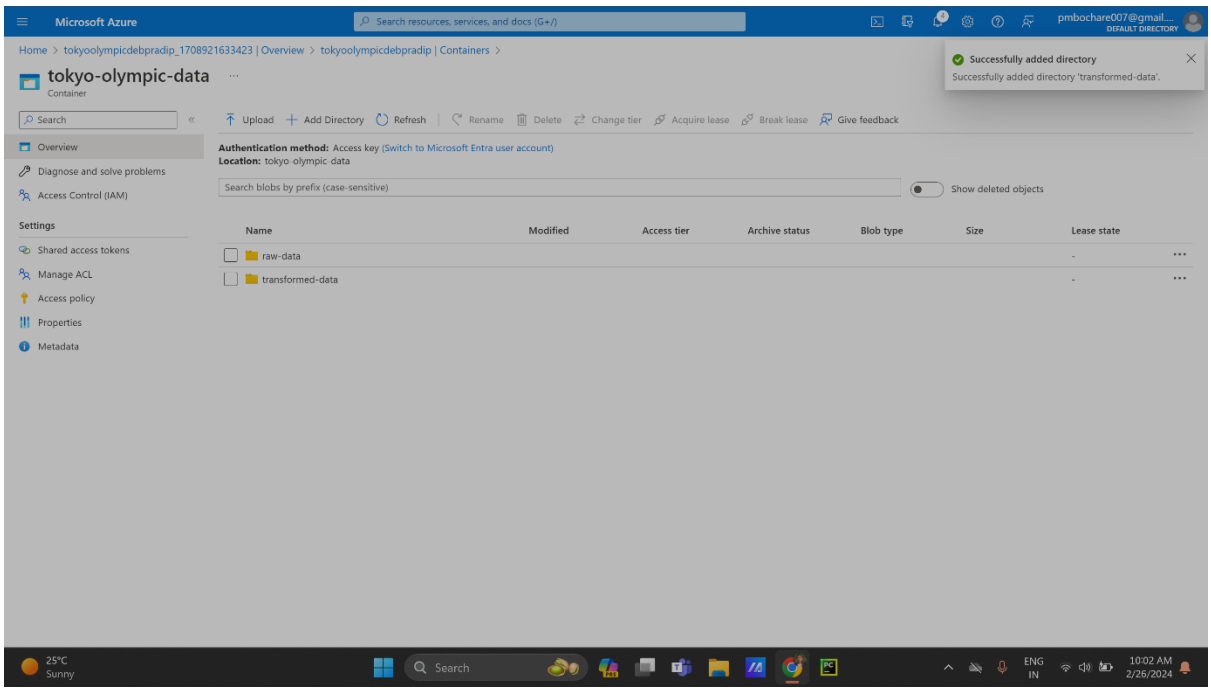
2. Create a Resource Group 'Tokyo-Olympic-data' to house and manage all the Azure resources associated with this project.

3. Within the created resource group, set up a storage account. This is specifically configured to leverage Azure Data Lake Storage (ADLS) Gen2 capabilities.





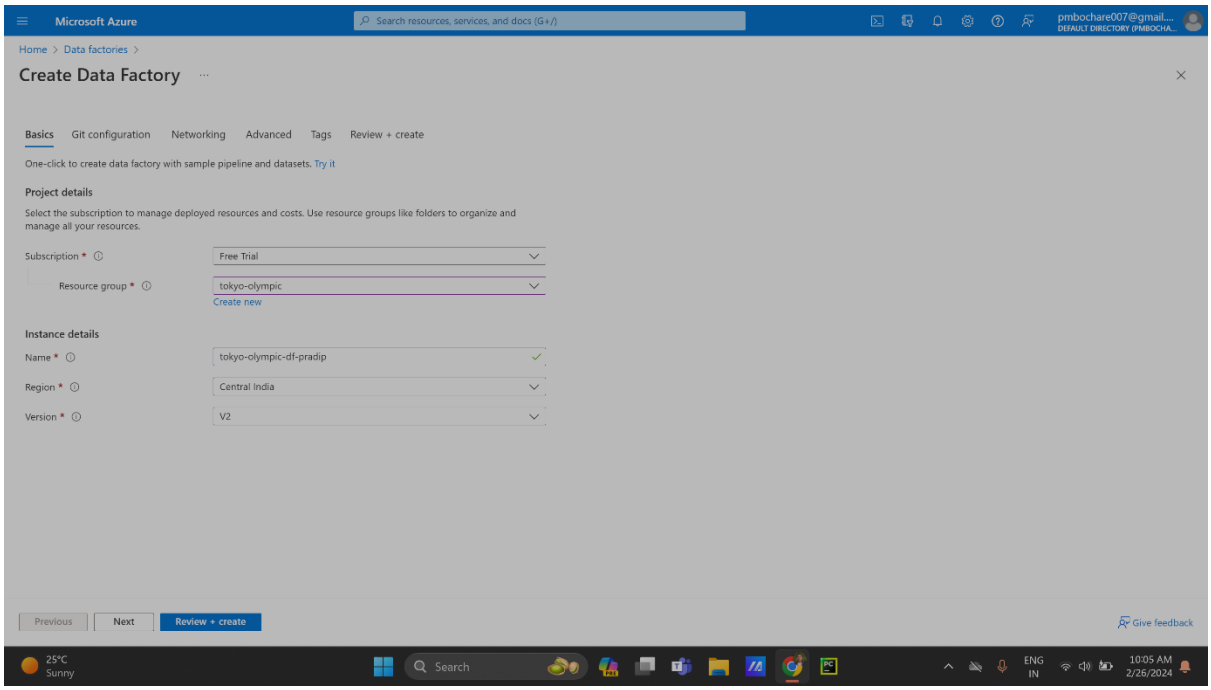
4. Create a Container inside this storage account to hold the project's data. Two directories 'raw-data' and 'transformed-data' are created to store raw data and transformed data. I made the two columns one for raw data integration and the other for the transformed data

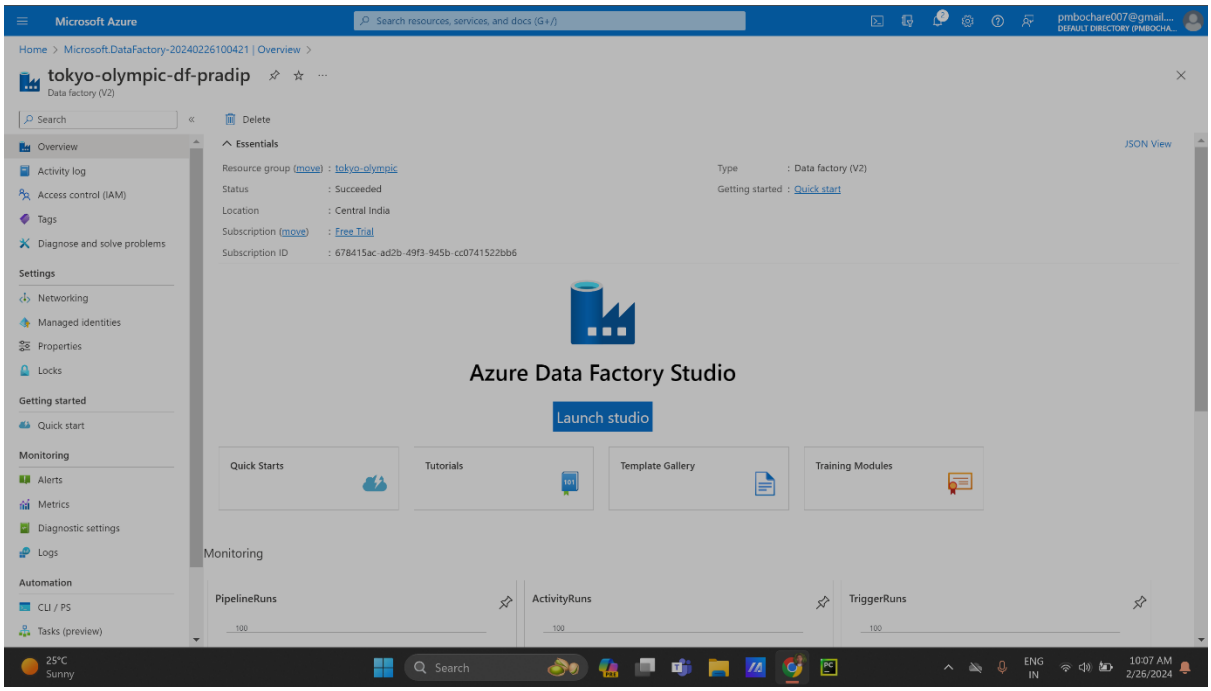


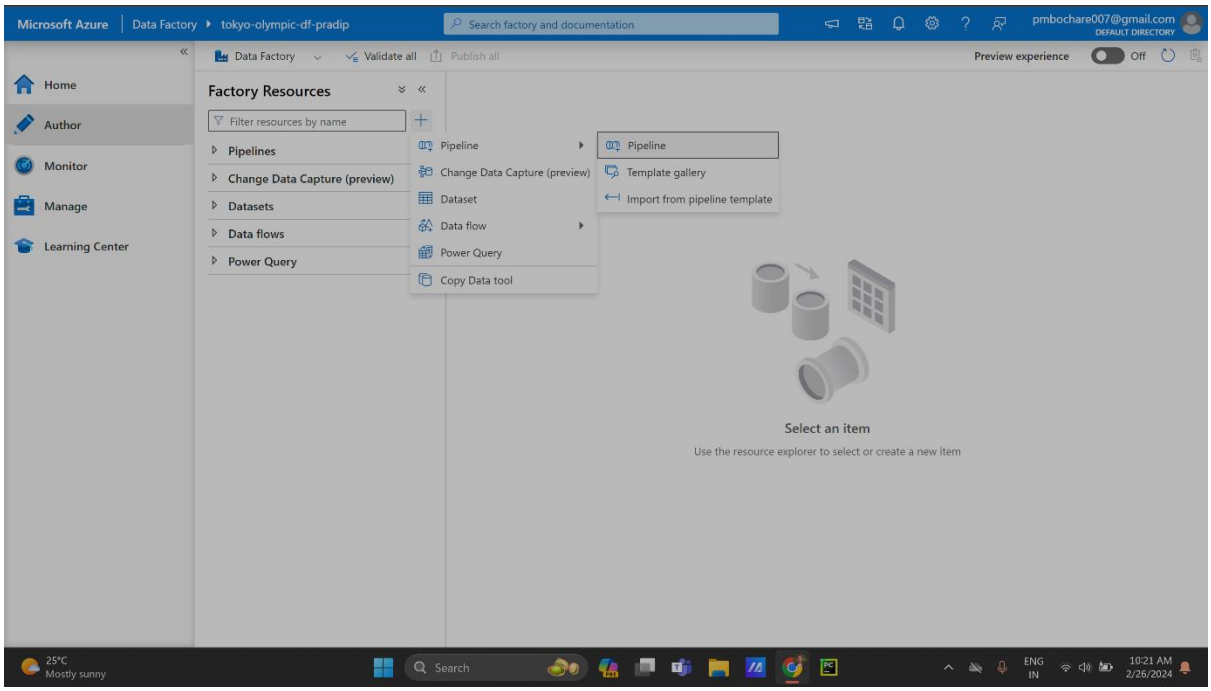
Data Ingestion using Azure Data Factory

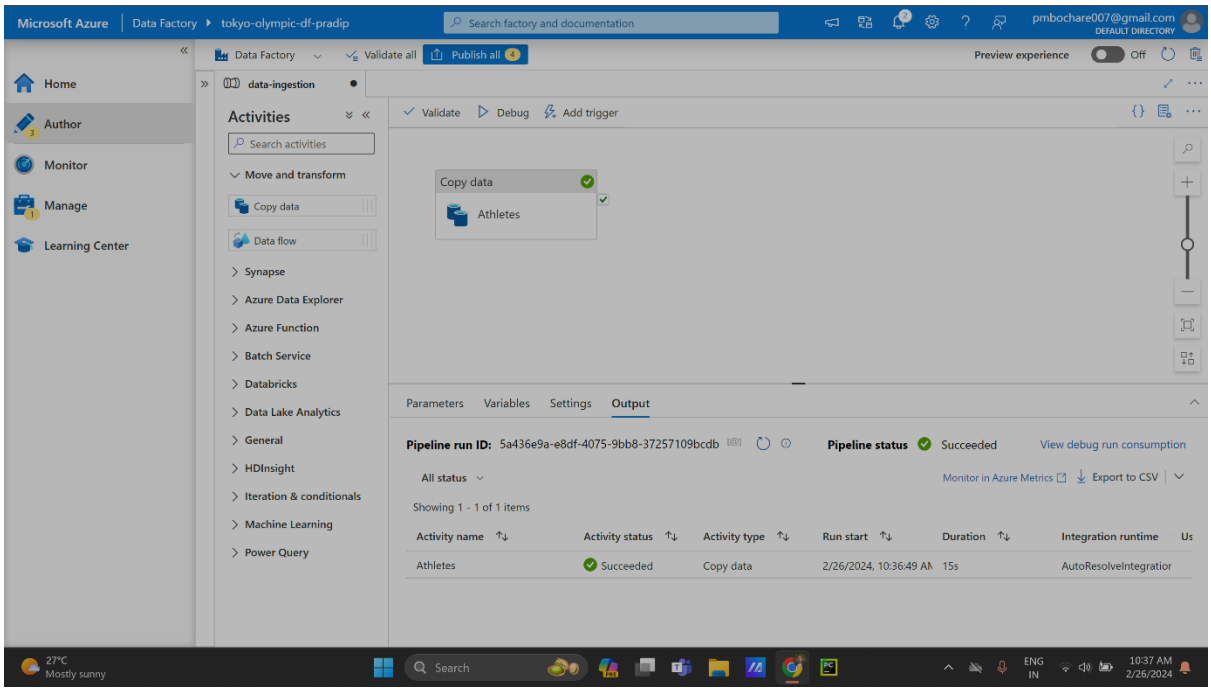
1. Begin by creating an Azure Data Factory workspace within the previously established resource group.
2. After setting up the workspace, launch the Azure Data Factory Studio.
3. Upload the Tokyo Olympics dataset from Kaggle to GitHub.
4. Within the studio, initialize a new data integration pipeline. Now use the task Copy Data to move data efficiently between various supported sources and destinations.
5. Configuring the Data Source with HTTP template as we are using http request to get the data from GitHub repo.
6. Establishing the Linked Service for source.
7. Configuring the File Format for and setting up the Linked Service Sink.
8. Repeat above steps to load all the datasets.
9. You can connect all the copy data activity together and run them all at once.

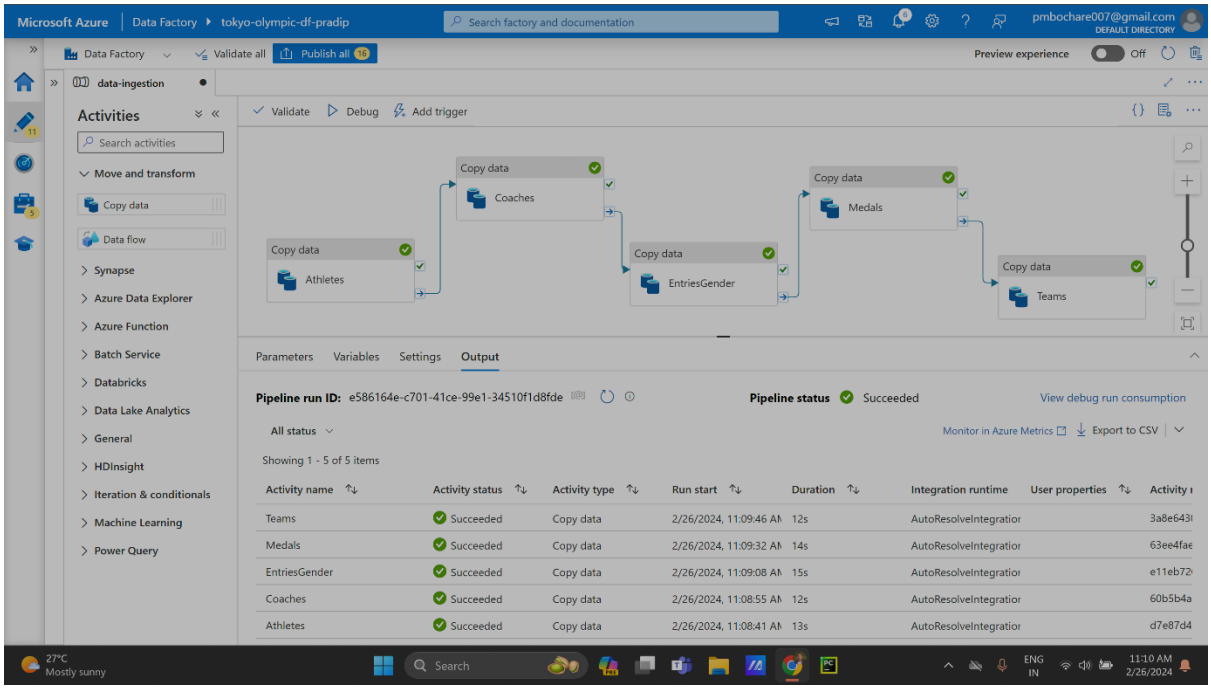
Creating Azure Data Factory



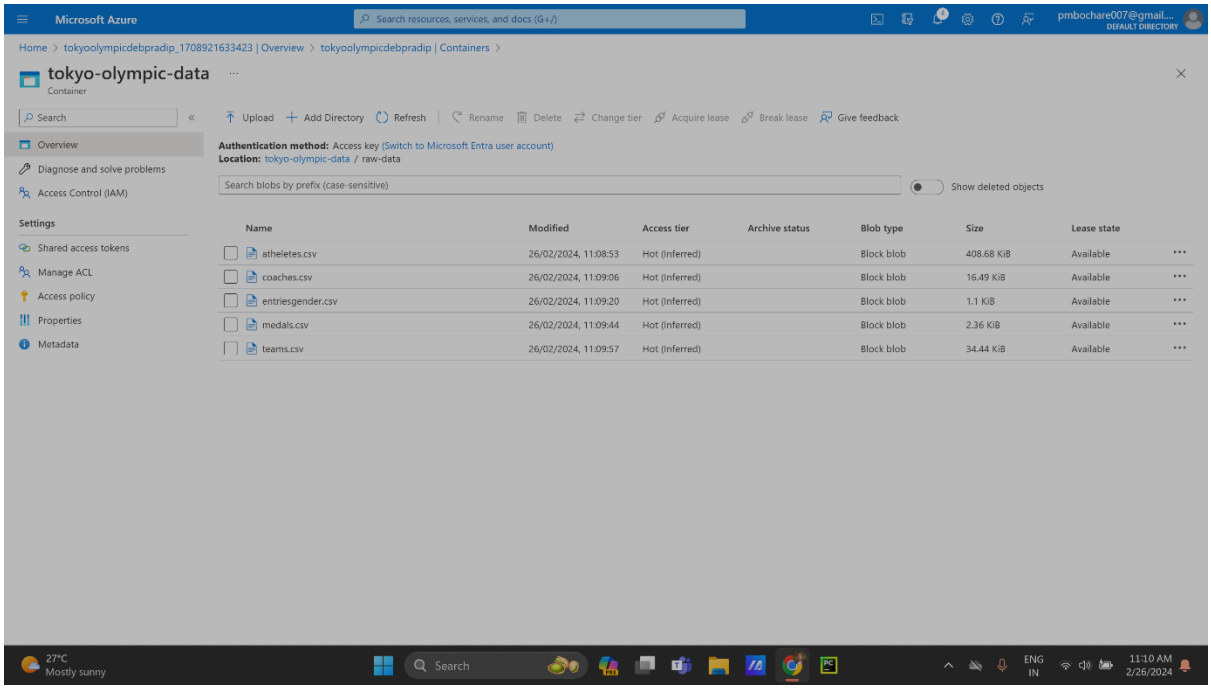



Microsoft Azure
Data Factory
tokyo-olympic-df-pradip
Search factory and documentation
pmbochare007@gmail.com
DEFAULT DIRECTORY
Data Factory
Validate all
Publish all
Preview experience
Off
Home
Author
Monitor
Manage
Learning Center
Factory Resources
Filter resources by name
Pipelines
Change Data Capture (preview)
Datasets
Data flows
Power Query
Pipeline
Change Data Capture (preview)
Dataset
Data flow
Power Query
Copy Data tool
Pipeline
Template gallery
Import from pipeline template
Select an item
Use the resource explorer to select or create a new item
25°C
Mostly sunny
Search
ENG
IN
10:21 AM
2/26/2024


Microsoft Azure
Data Factory
tokyo-olympic-df-pradip
Search factory and documentation
pmbochare007@gmail.com
DEFAULT DIRECTORY
Data Factory
Validate all
Publish all 4
Preview experience
Off
Home
Author
Monitor
Manage
Learning Center
data-ingestion
Activities
Search activities
Move and transform
Copy data
Data flow
Synapse
Azure Data Explorer
Azure Function
Batch Service
Databricks
Data Lake Analytics
General
HDInsight
Iteration & conditionals
Machine Learning
Power Query
Validate
Debug
Add trigger
Copy data
Athletes
Parameters
Variables
Settings
Output
Pipeline run ID: 5a436e9a-e8df-4075-9bb8-37257109bcdb
Pipeline status Succeeded
View debug run consumption
All status
Monitor in Azure Metrics
Export to CSV
Showing 1 - 1 of 1 items
Activity name
Activity status
Activity type
Run start
Duration
Integration runtime
Athletes
Succeeded
Copy data
2/26/2024, 10:36:49 AM
15s
AutoResolveIntegratior
27°C
Mostly sunny
Search
ENG
IN
10:37 AM
2/26/2024



10. After the pipeline completes its execution, navigate to your Azure Data Lake Storage Gen2. Dive into the "raw data" folder and validate that the files, like "athletes.csv", "medals.csv", etc., are present and populated with the expected data.

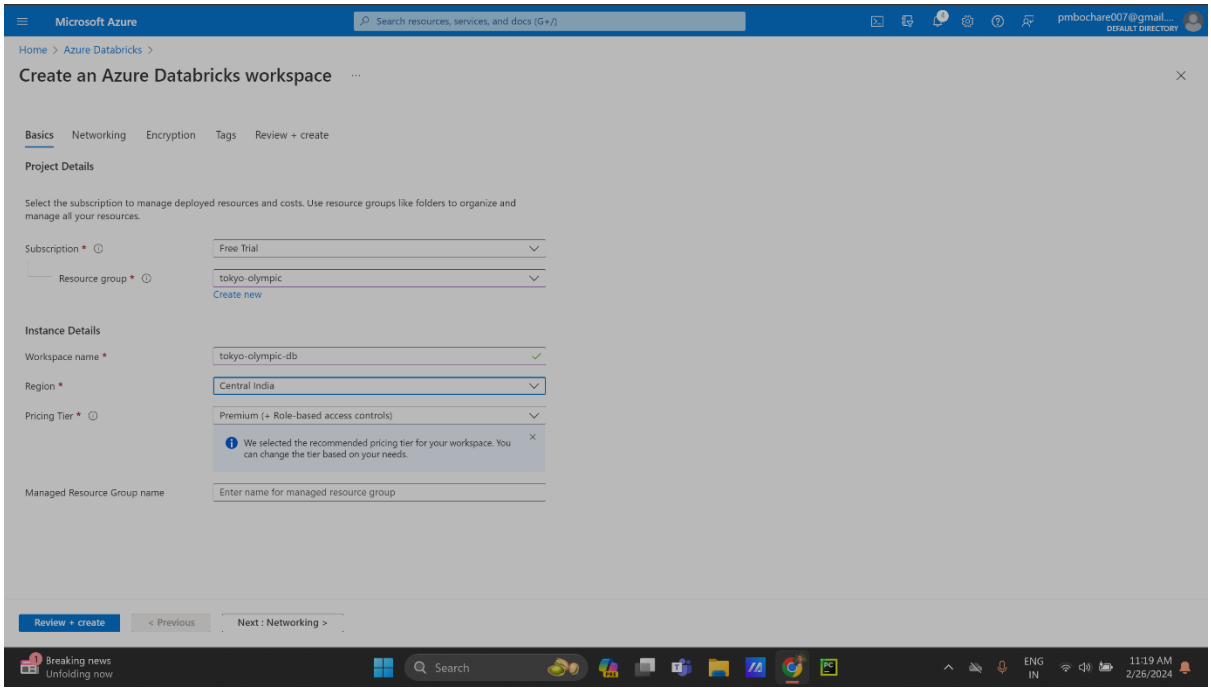


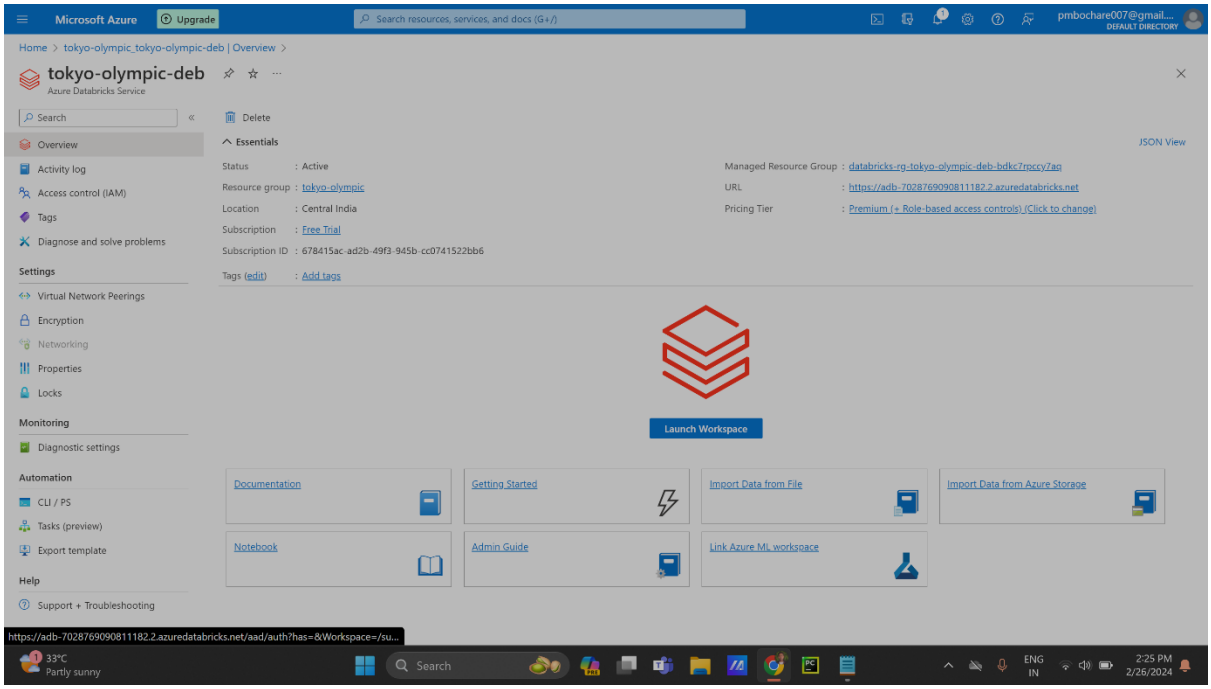
Data Transformation using Azure Databricks

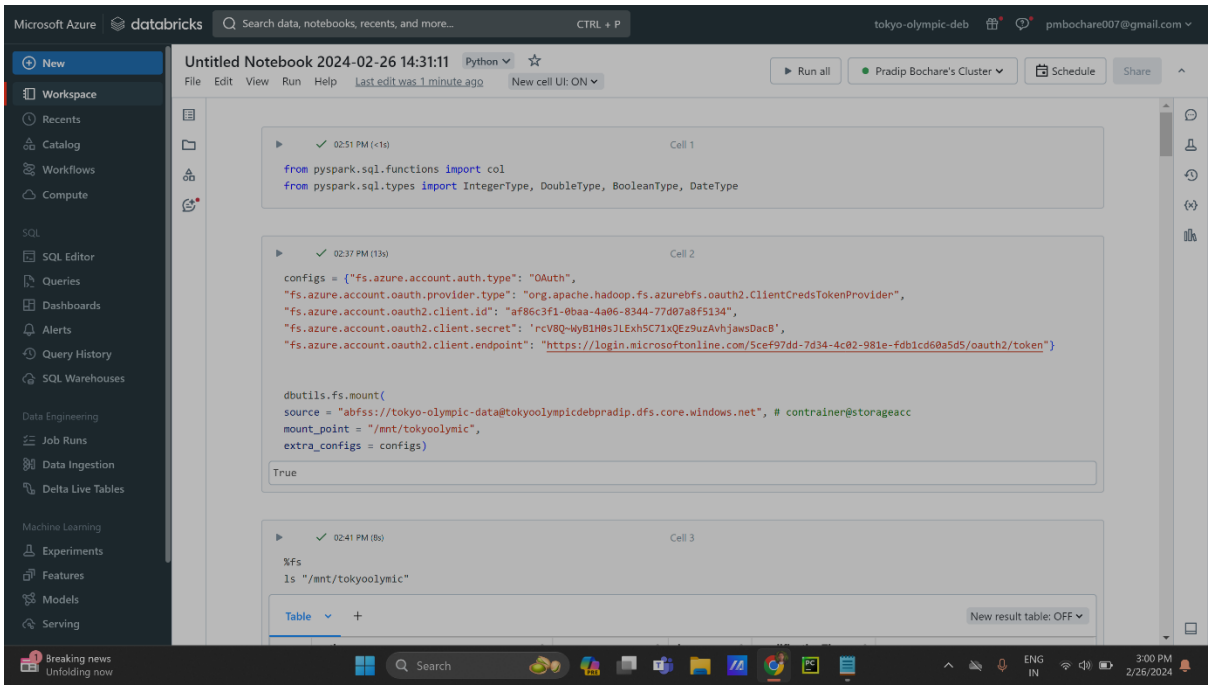
1. Navigate to Azure Databricks within the Azure portal and create a workspace within the previously established resource group and launch it.
2. Configuring Compute in Databricks
3. Create a new notebook within Databricks and rename it appropriately, reflecting its purpose or the dataset it pertains to.
4. Establishing a Connection to Azure Data Lake Storage (ADLS)
5. Using the credentials (Client ID, Tenant ID, Secret), write the appropriate code in the Databricks notebook to mount ADLS.
6. Writing Data Transformations mount ADLS Gen2 to Databricks.
7. Writing Transformed Data to ADLS Gen2.

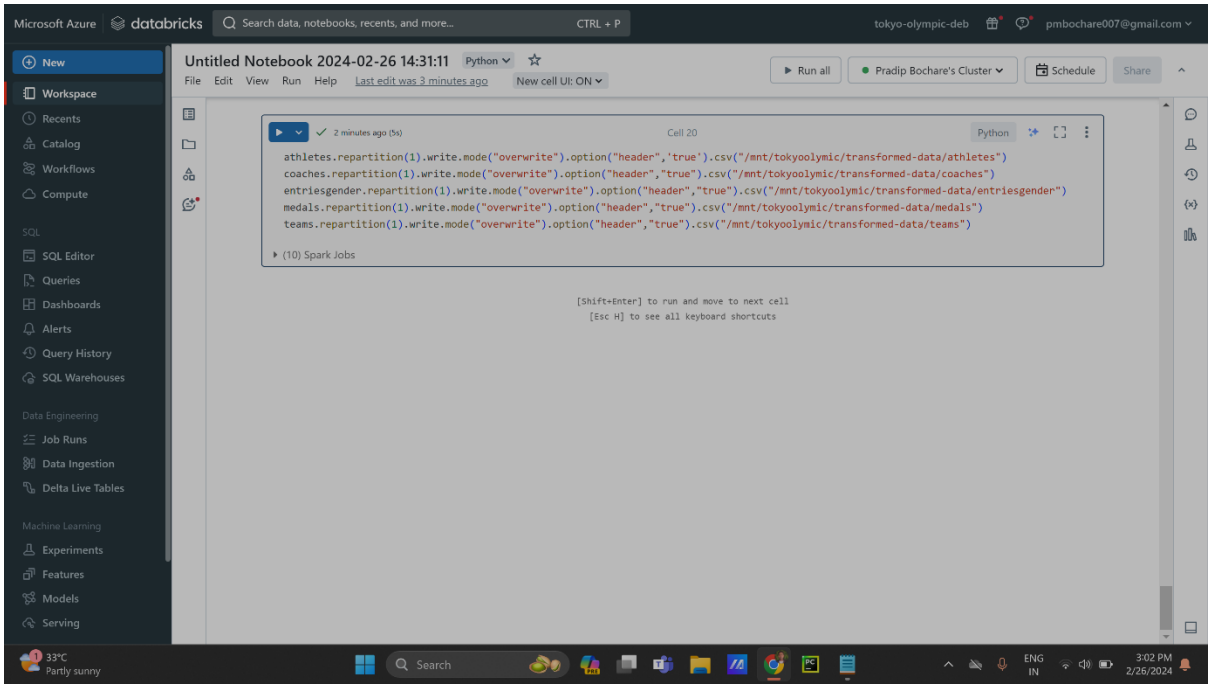
After integration I connected my data lake to Azure Data Bricks using secret key. There is used Apache spark code for data cleaning like fixing the data types of various columns

Creating Azure Databricks

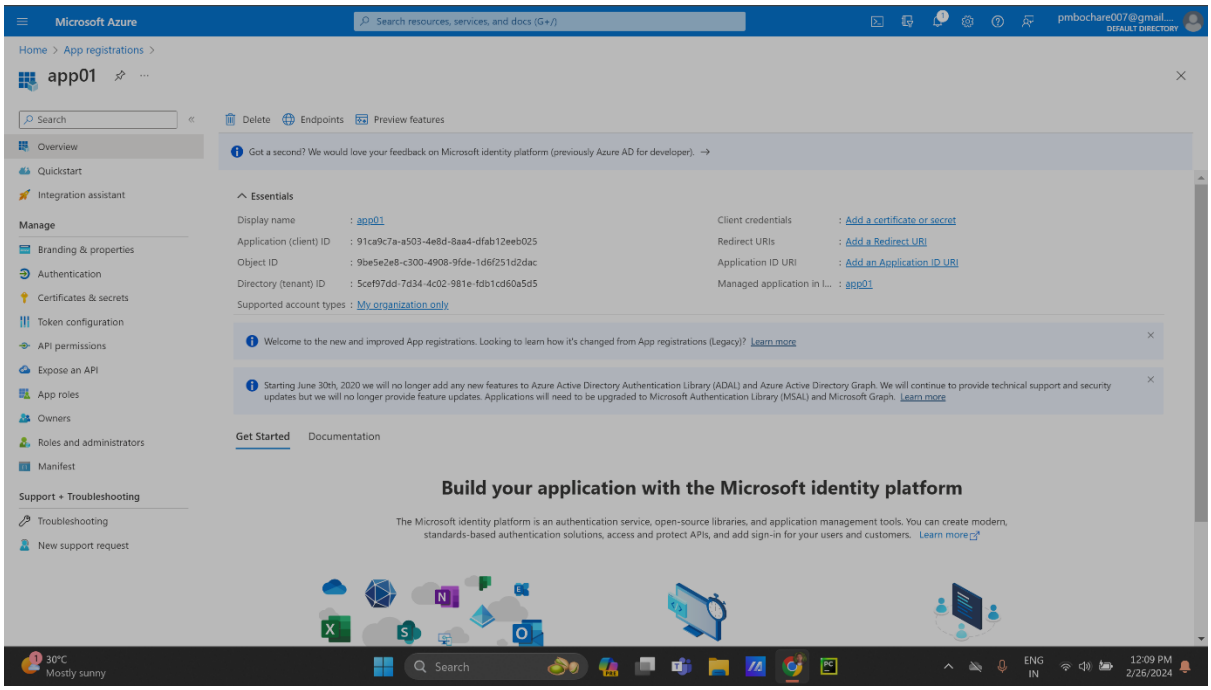



Microsoft Azure
Upgrade
Search resources, services, and docs (G+/)
pmbochare007@gmail....
DEFAULT DIRECTORY
Home > tokyo-olympic_tokyo-olympic-deb | Overview >
tokyo-olympic-deb
Azure Databricks Service
Search
Overview
Activity log
Access control (IAM)
Tags
Diagnose and solve problems
Settings
Virtual Network Peerings
Encryption
Networking
Properties
Locks
Monitoring
Diagnostic settings
Automation
CLI / PS
Tasks (preview)
Export template
Help
Support + Troubleshooting
Delete
Essentials
JSON View
Status : Active
Resource group : tokyo-olympic
Location : Central India
Subscription : Free Trial
Subscription ID : 678415ac-ad2b-49f3-945b-cc0741522bb6
Tags (edit) : Add tags
Managed Resource Group : databricks-rg-tokyo-olympic-deb-bdkc7rpccy7aq
URL : https://adb-7028769090811182.2.azuredatabricks.net
Pricing Tier : Premium (+ Role-based access controls) (Click to change)
Launch Workspace
Documentation
Getting Started
Import Data from File
Import Data from Azure Storage
Notebook
Admin Guide
Link Azure ML workspace
https://adb-7028769090811182.2.azuredatabricks.net/aad/auth?has=&Workspace=/su...
33°C
Partly sunny
Search
ENG
IN
2:25 PM
2/26/2024


Microsoft Azure
databricks
Search data, notebooks, recents, and more...
CTRL + P
tokyo-olympic-deb
pmbochare007@gmail.com
New
Workspace
Recents
Catalog
Workflows
Compute
SQL
SQL Editor
Queries
Dashboards
Alerts
Query History
SQL Warehouses
Data Engineering
Job Runs
Data Ingestion
Delta Live Tables
Machine Learning
Experiments
Features
Models
Serving
Untitled Notebook 2024-02-26 14:31:11
Python
File Edit View Run Help
Last edit was 1 minute ago
New cell UI: ON
Run all
Pradip Bochare's Cluster
Schedule
Share
02:51 PM (<1s)
Cell 1
from pyspark.sql.functions import col
from pyspark.sql.types import IntegerType, DoubleType, BooleanType, DateType
02:37 PM (13s)
Cell 2
configs = {"fs.azure.account.auth.type": "OAuth",
"fs.azure.account.oauth.provider.type": "org.apache.hadoop.fs.azurebfs.oauth2.ClientCredsTokenProvider",
"fs.azure.account.oauth2.client.id": "af86c3f1-0baa-4a06-8344-77d07a8f5134",
"fs.azure.account.oauth2.client.secret": 'rcV8Q~WyB1H0sJLExh5C71xQEz9uzAvhjawsDacB',
"fs.azure.account.oauth2.client.endpoint": "https://login.microsoftonline.com/5cef97dd-7d34-4c02-981e-fdb1cd60a5d5/oauth2/token"}
dbutils.fs.mount(
source = "abfss://tokyo-olympic-data@tokyoolympicdebpradip.dfs.core.windows.net", # contrainer@storageacc
mount_point = "/mnt/tokyoolymic",
extra_configs = configs)
True
02:41 PM (8s)
Cell 3
%fs
ls "/mnt/tokyoolymic"
Table
New result table: OFF
Breaking news
Unfolding now
Search
ENG
IN
3:00 PM
2/26/2024



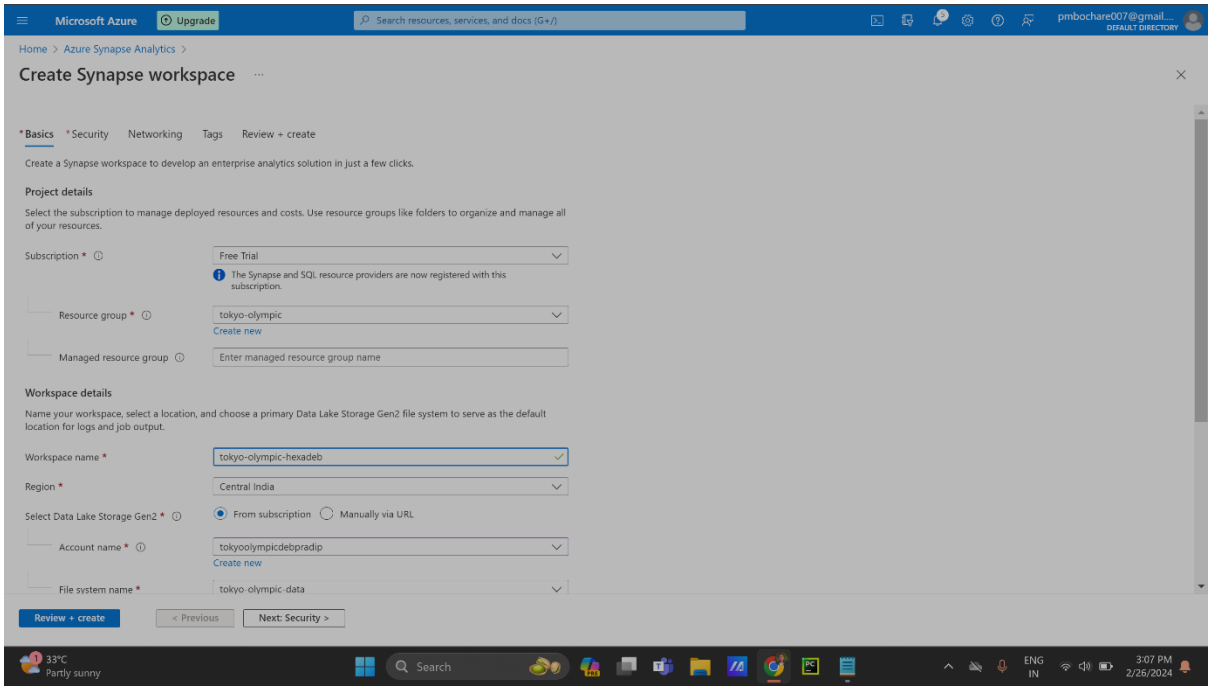
After integration I connected my data lake to Azure Data Bricks using secret key. There is used Apache spark code for data cleaning like fixing the data types of various columns.

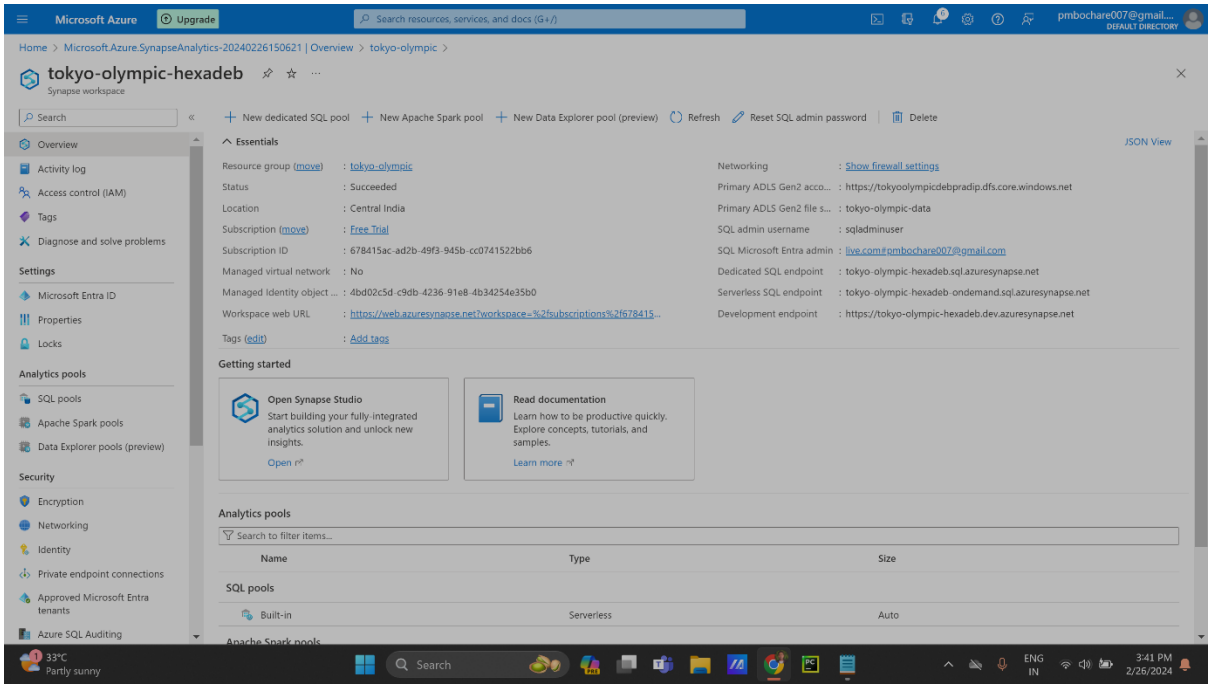


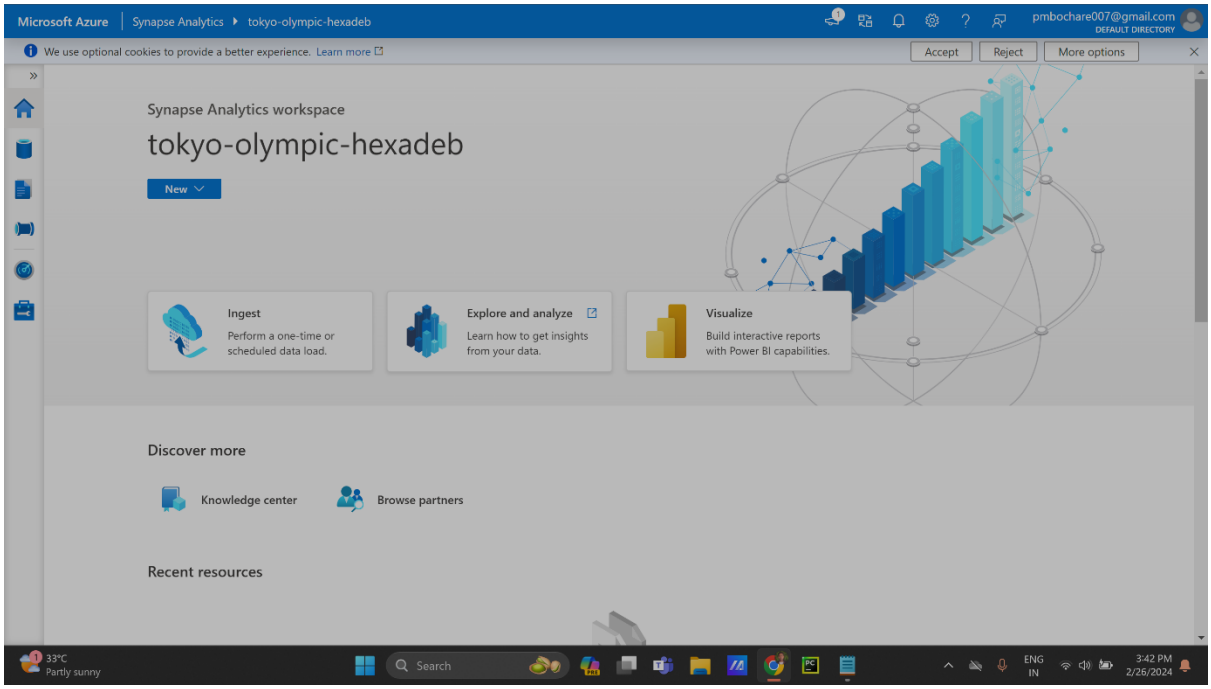
Setting Up and Using Azure Synapse Analytics

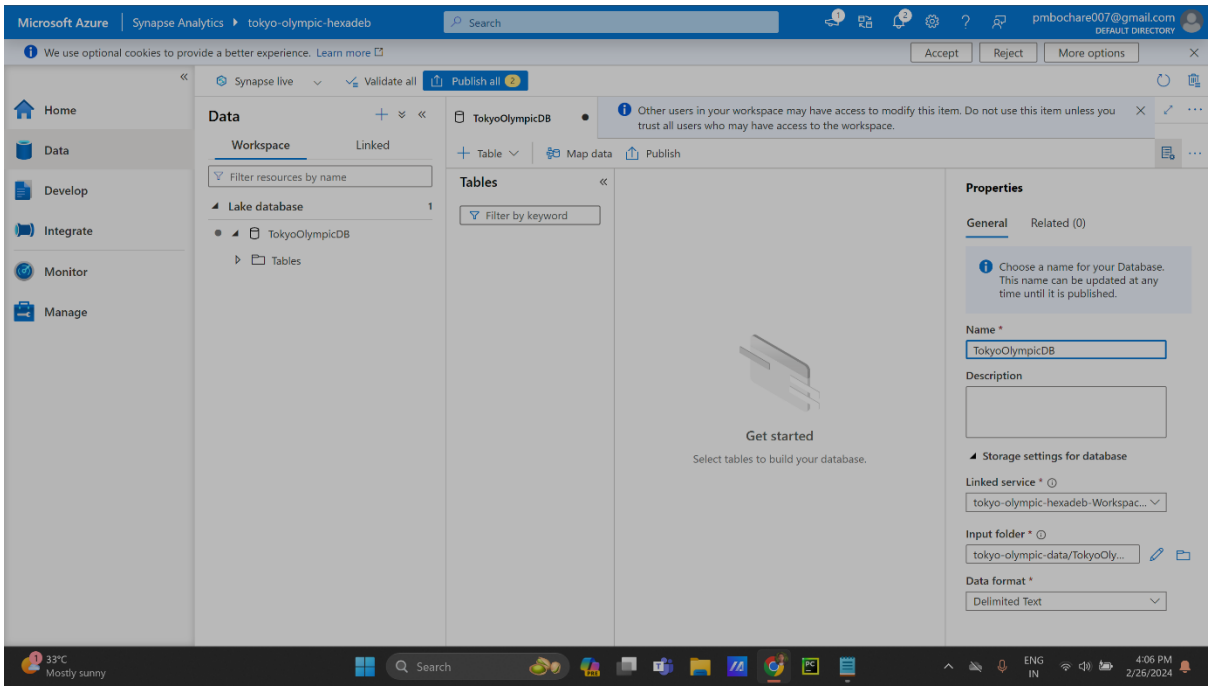
1. Creating a Synapse Analytics Workspace.
2. Within Workspace navigate to the "Data" section, choose "Lake Database" and create a Database "TokyoOlympicDB"
3. Creating Table from Data Lake from the Transformed Data folder within your ADLS Gen2 storage.

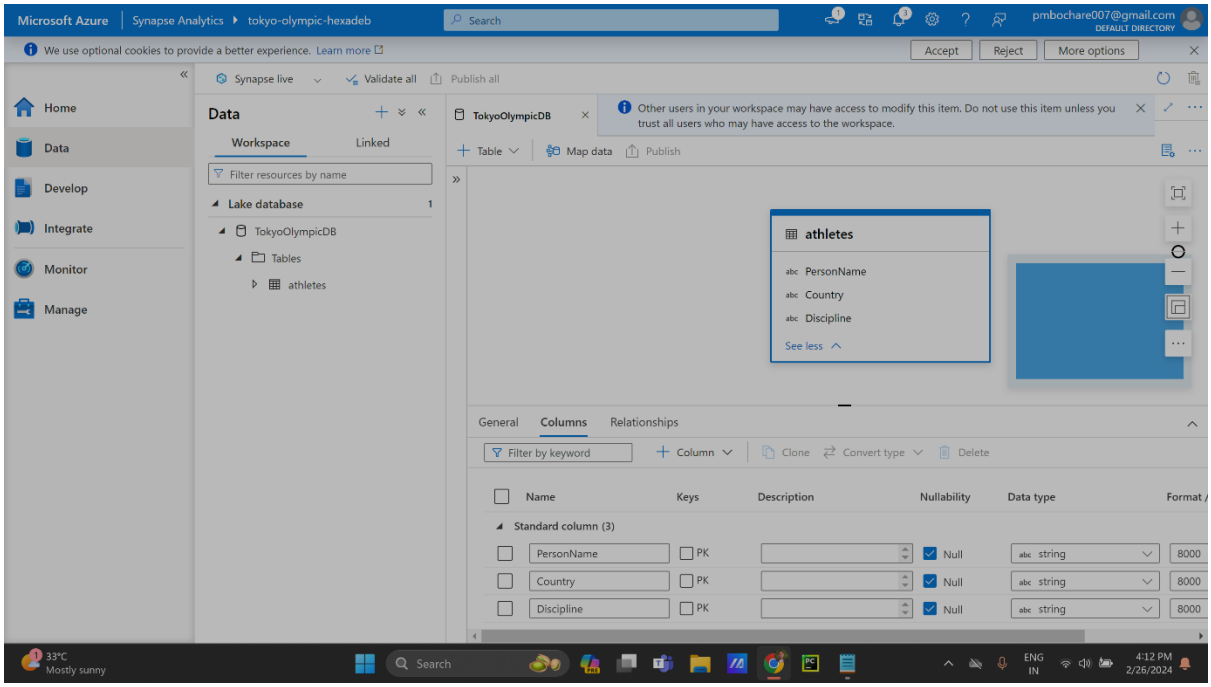
Creating Azure synapse Analytics

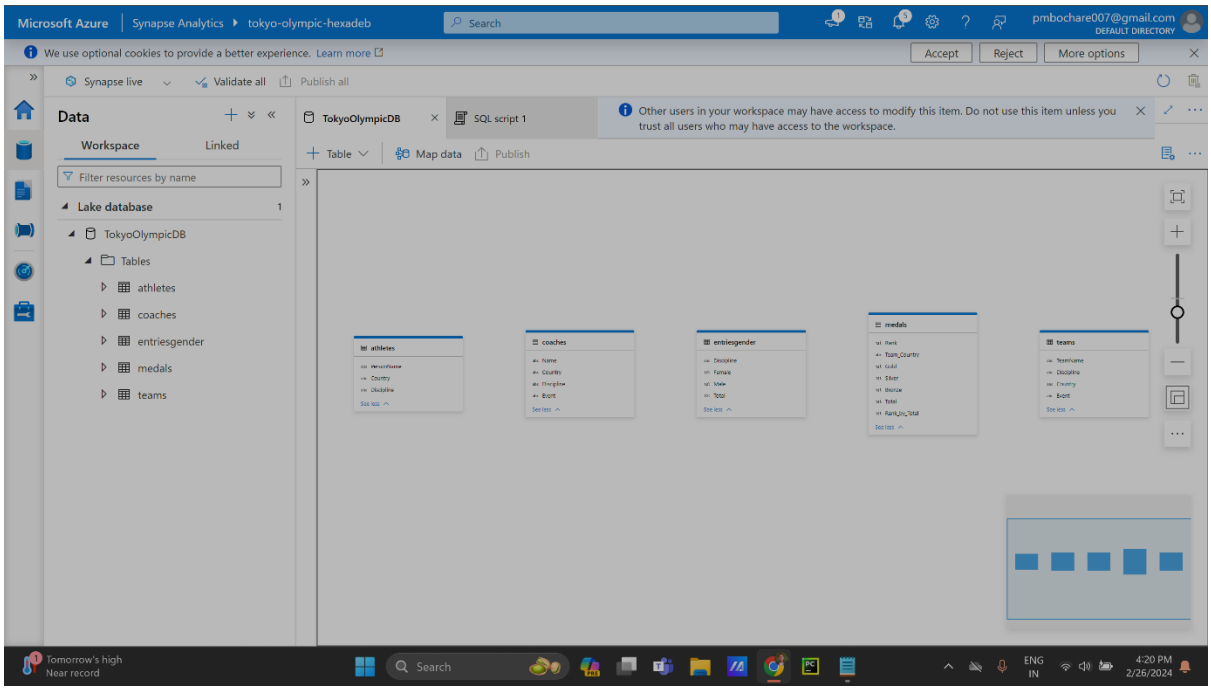




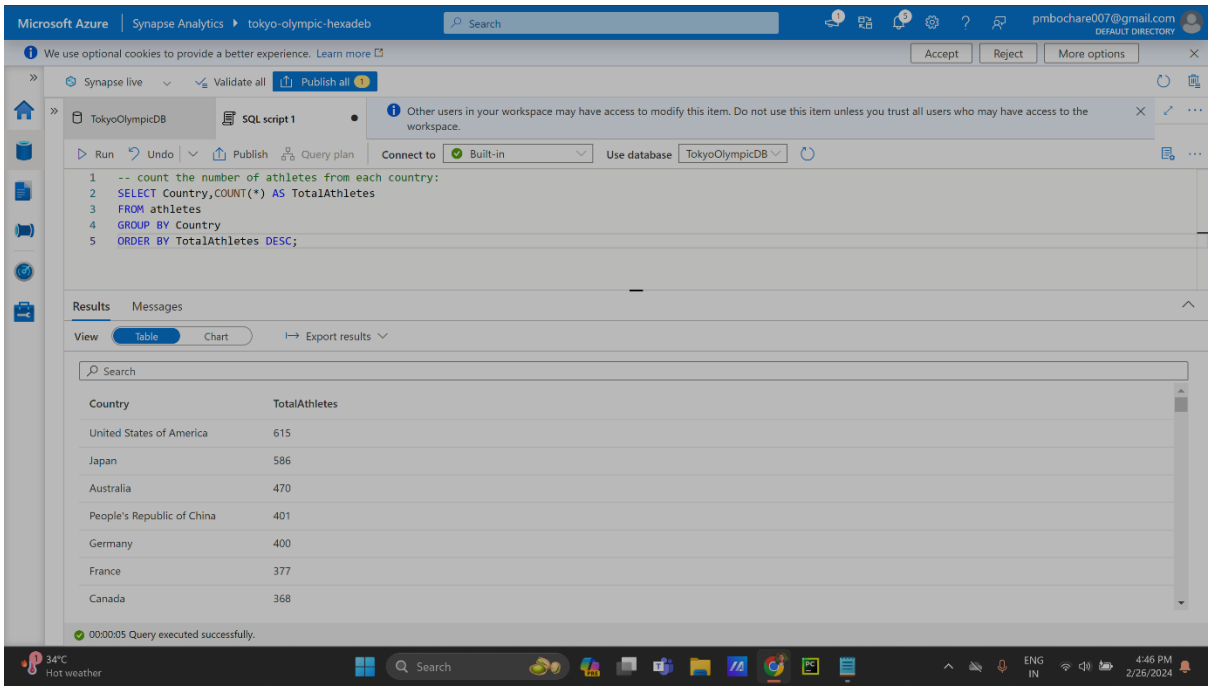

Microsoft Azure
Synapse Analytics
tokyo-olympic-hexadeb
pmbochare007@gmail.com
DEFAULT DIRECTORY
We use optional cookies to provide a better experience. Learn more
Accept
Reject
More options
Synapse Analytics workspace
tokyo-olympic-hexadeb
New
Ingest
Perform a one-time or scheduled data load.
Explore and analyze
Learn how to get insights from your data.
Visualize
Build interactive reports with Power BI capabilities.
Discover more
Knowledge center
Browse partners
Recent resources
33°C
Partly sunny
Search
ENG
IN
3:42 PM
2/26/2024

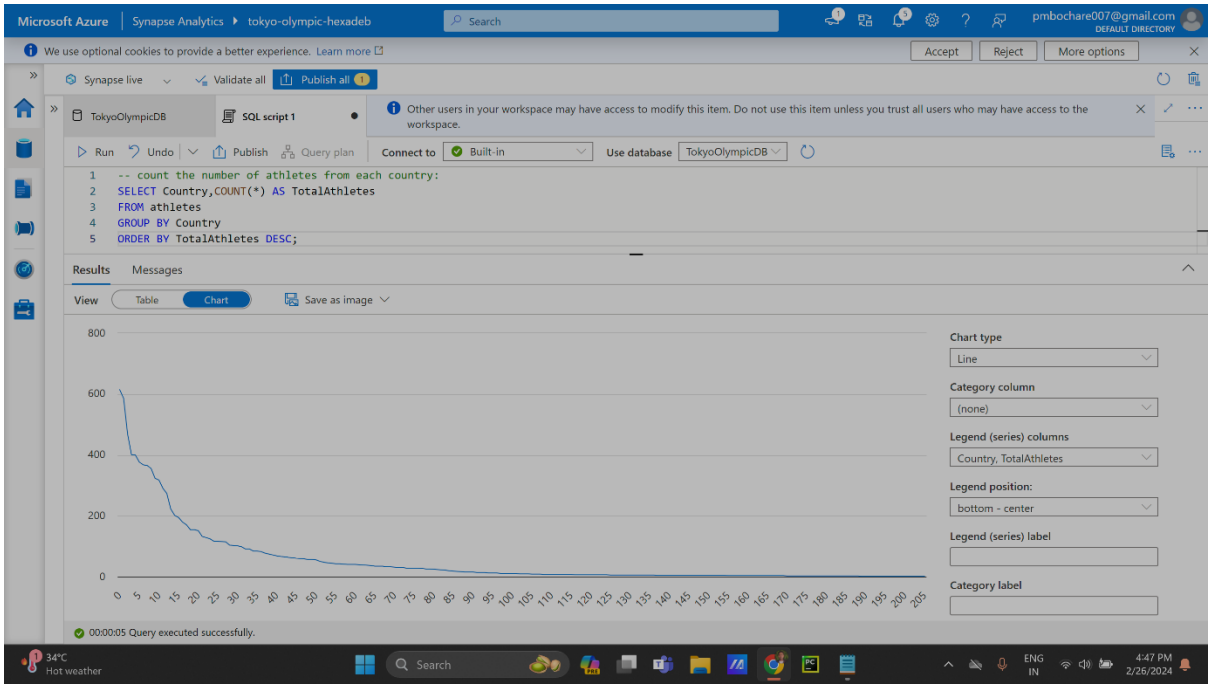

Microsoft Azure
Synapse Analytics
tokyo-olympic-hexadeb
Search
pmbochare007@gmail.com
DEFAULT DIRECTORY
We use optional cookies to provide a better experience. Learn more
Accept
Reject
More options
Synapse live
Validate all
Publish all
Home
Data
Develop
Integrate
Monitor
Manage
Data
Workspace
Linked
Filter resources by name
Lake database
TokyoOlympicDB
Tables
TokyoOlympicDB
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.
Table
Map data
Publish
Tables
Filter by keyword
Get started
Select tables to build your database.
Properties
General
Related (0)
Choose a name for your Database. This name can be updated at any time until it is published.
Name *
TokyoOlympicDB
Description
Storage settings for database
Linked service *
tokyo-olympic-hexadeb-Workspac...
Input folder *
tokyo-olympic-data/TokyoOly...
Data format *
Delimited Text
33°C
Mostly sunny
Search
ENG
IN
4:06 PM
2/26/2024


Microsoft Azure
Synapse Analytics
tokyo-olympic-hexadeb
Search
pmbochare007@gmail.com
DEFAULT DIRECTORY
We use optional cookies to provide a better experience. Learn more
Accept
Reject
More options
Synapse live
Validate all
Publish all
Home
Data
Develop
Integrate
Monitor
Manage
Data
Workspace
Linked
Filter resources by name
Lake database
TokyoOlympicDB
Tables
athletes
TokyoOlympicDB
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.
Table
Map data
Publish
athletes
PersonName
Country
Discipline
See less
General
Columns
Relationships
Filter by keyword
Column
Clone
Convert type
Delete
Name
Keys
Description
Nullability
Data type
Format
Standard column (3)
PersonName
PK
Null
string
8000
Country
PK
Null
string
8000
Discipline
PK
Null
string
8000
33°C
Mostly sunny
Search
ENG
IN
4:12 PM
2/26/2024


Microsoft Azure
Synapse Analytics
tokyo-olympic-hexadeb
Search
pmbochare007@gmail.com
DEFAULT DIRECTORY
We use optional cookies to provide a better experience. Learn more
Accept
Reject
More options
Synapse live
Validate all
Publish all
Data
Workspace
Linked
Filter resources by name
Lake database
TokyoOlympicDB
Tables
athletes
coaches
entriesgender
medals
teams
TokyoOlympicDB
SQL script 1
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.
Table
Map data
Publish
Tomorrow's high
Near record
Search
ENG
IN
4:20 PM
2/26/2024

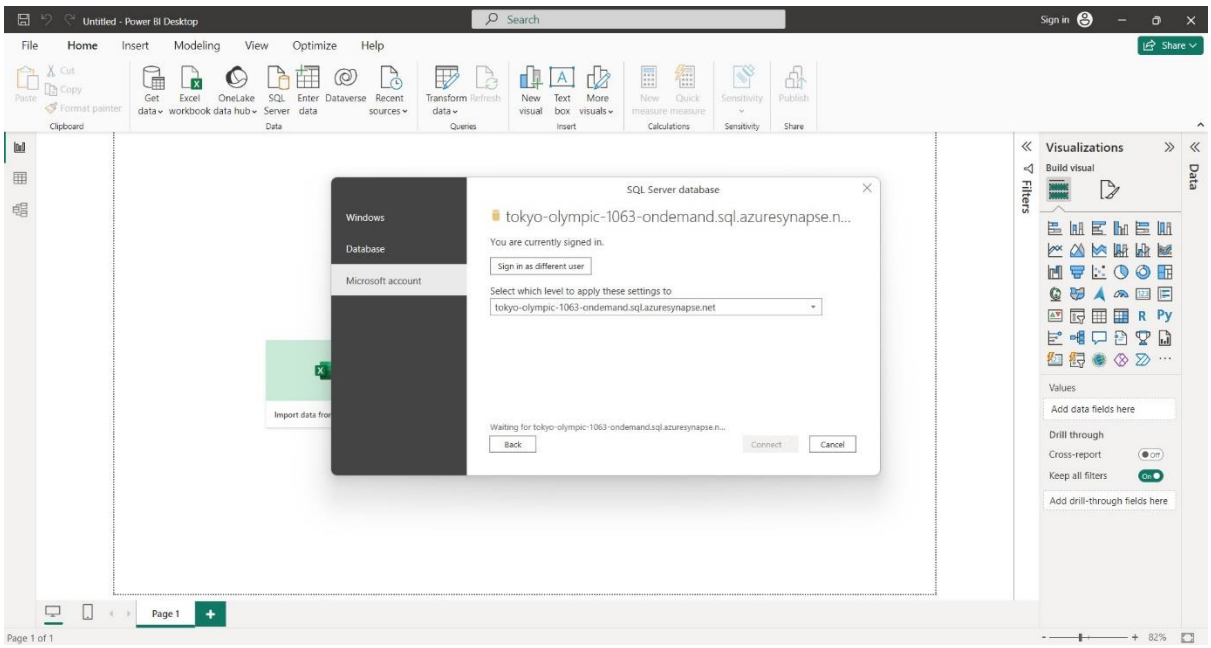
Performing Analysis using SQL scripts and charts

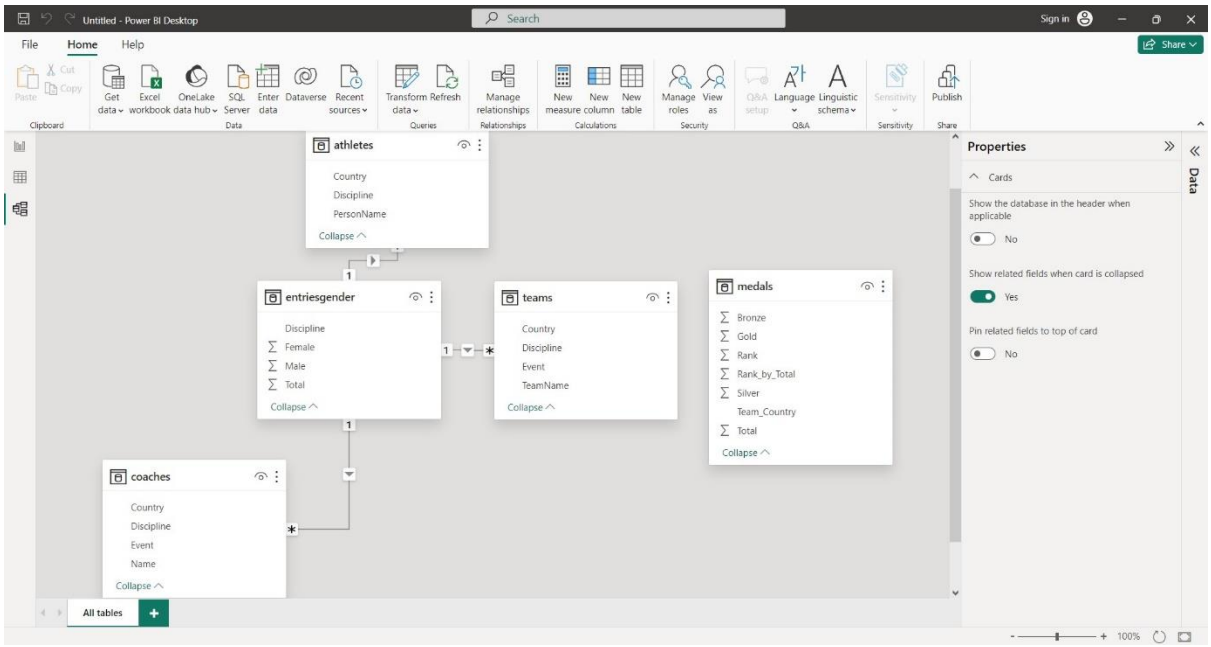


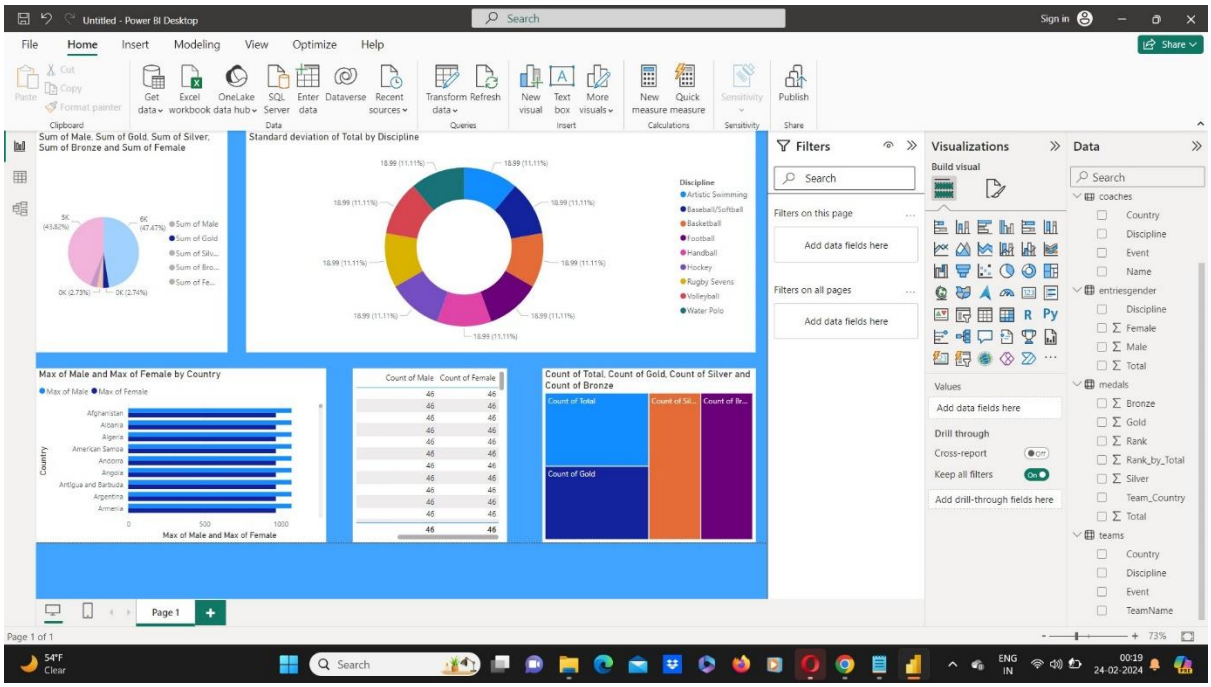


Performing Data Analysis on the Data using PowerBI

After that I connect to the PowerBI to visualize the Tokyo-Olympic data. owner BI allows users to create interactive and visually appealing reports and dashboards from their data. Users can choose from a variety of visualization options such as charts, graphs, maps, and tables to represent their data in meaningful ways



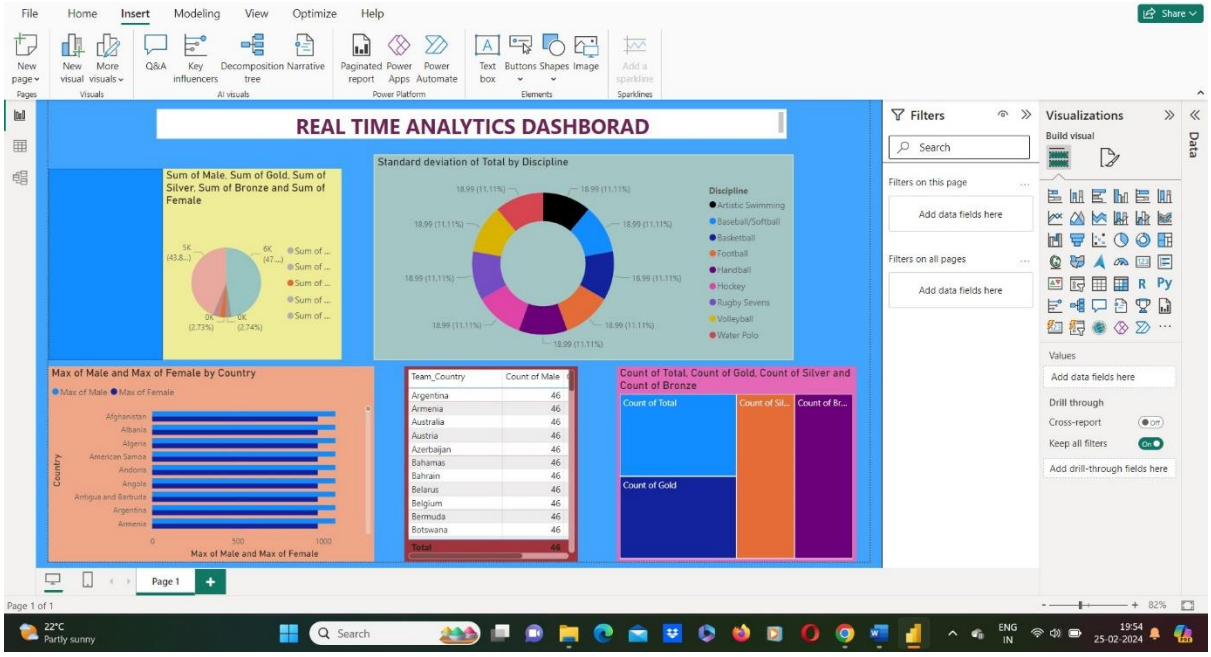




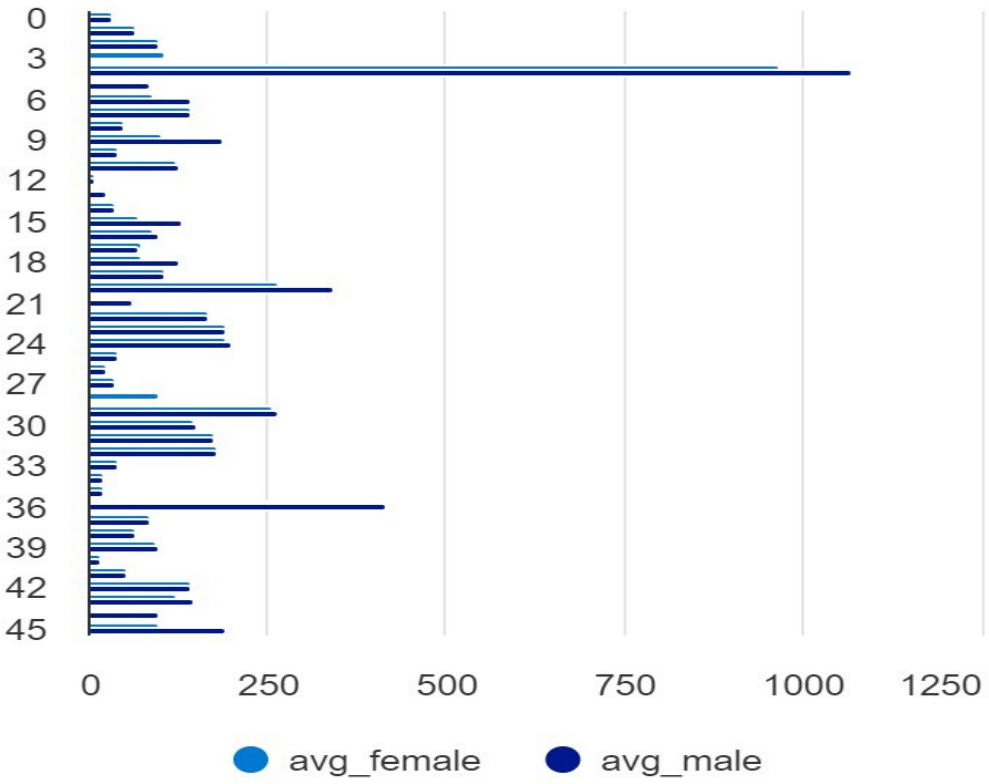
6.1 FINAL OUTPUT SCREENSHOT

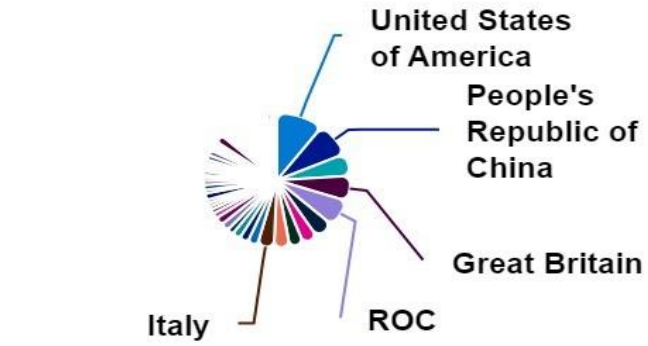
The final output screenshot of the analytics dashboard is:

The final output is an interactive dashboard created using Power BI, providing users with real-time insights into their data. The dashboard includes various visualizations such as charts, graphs, tables, and maps to represent key metrics and trends. The final output adheres to data governance and security best practices to ensure data confidentiality, integrity, and availability.

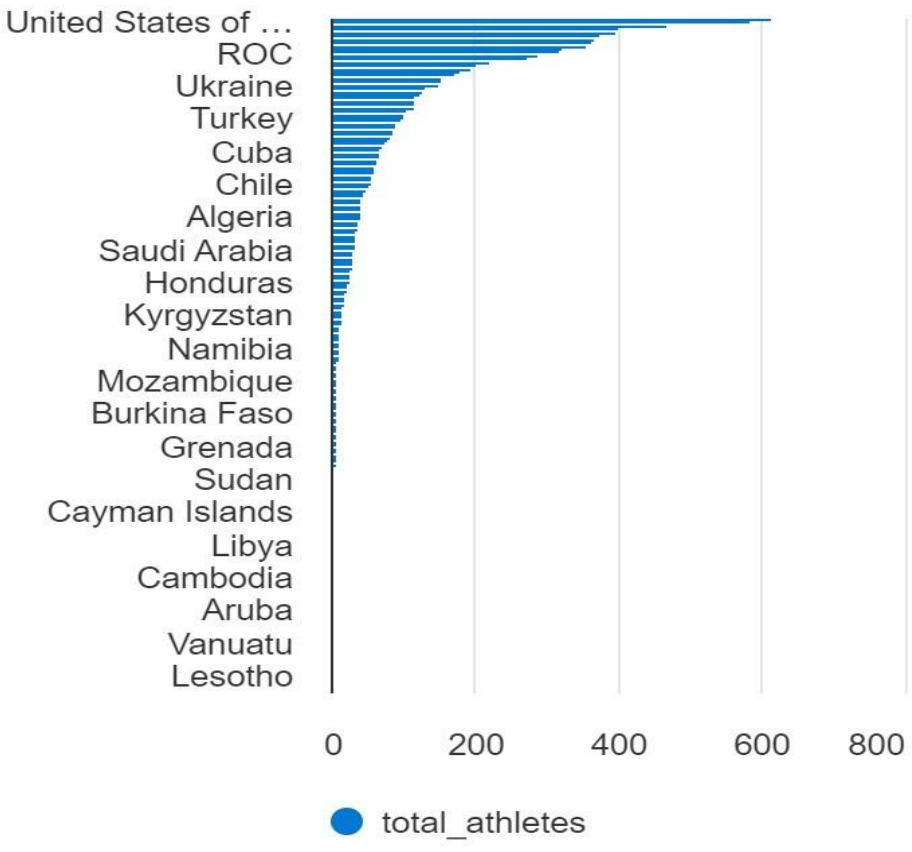


Some sample outputs of visualizations are:




United States of America
People's Republic of China
Great Britain
ROC
Italy


United States of America
People's Republic of China
Japan
Great Britain
ROC
Australia
1/16


United States of ...
ROC
Ukraine
Turkey
Cuba
Chile
Algeria
Saudi Arabia
Honduras
Kyrgyzstan
Namibia
Mozambique
Burkina Faso
Grenada
Sudan
Cayman Islands
Libya
Cambodia
Aruba
Vanuatu
Lesotho
0
200
400
600
800
total_athletes

CONCLUSIONS

The conclusion of the project to build real-time analytics is the project successfully achieved its objectives of creating a real-time analytics dashboard that provides insights into data stored in ADLS Gen2 and Azure Synapse Analytics. Leveraging Azure Databricks facilitated real-time analytics processing, enabling the extraction of valuable insights from streaming data. Azure Data Factory effectively orchestrated data pipelines, ensuring seamless and reliable data ingestion from various sources into ADLS Gen2 and Azure Synapse Analytics. Power BI empowered users with interactive and visually appealing dashboards, facilitating the interpretation of analytics insights and enabling data-driven decision-making. The real-time analytics dashboard provided stakeholders and decision-makers with timely and actionable insights, enabling them to respond promptly to changing business conditions and make informed decisions.