TECHNO INTERNATIONAL NEW TOWN

Block-DG 1/1, Action Area 1, New Town, Kolkata -700156, West Bengal, India

Department of Computer Science and Business System

Seventh Semester Project Evaluation-I Report (PCC-CSBS781)

Olympics Data Analysis using Microsoft Azure

Prepared by

Debajit Samanta(18731120017)

Subhodeep Roy(18731120002)

Under the Guidance of

Prof. Bitan Misra

Batch:- 2020-2024 Semester :7 th (2023 –ODD) Year : July 2023 – December 2023

Stream:- Computer Science and Business System Year of Study: 4th

Affiliated to

MAULANA ABUL KALAM AZAD
UNIVERSITY OF TECHNOLOGY,
WEST BENGAL

MAULANA ABUL KALAM AZAD UNIVERSITY OF TECHNOLOGY, WESTBENGAL

(FORMERLY KNOWN AS WEST BENGAL UNIVERSITY OF TECHNOLOGY)

ACKNOWLEDGEMENT

We would like to express our sincere gratitude to Prof. Bitan Misra of the Department of CSBS / CSE, whose role as project guide was invaluable for the project. We are extremely thankful for the keen interest she took in advising us, for the books and reference materials provided for the moral support extended to us.

Last but not the least we convey our gratitude to all the teachers for providing us the technical skill that will always remain as our asset and to all non-teaching staff for the cordial support they offered.

Place: Techno International New Town

Date: 04/12/2023

Debajit Samanta
(Roll No: - 18731120017)

Subhodeep Roy
(Roll No: - 18731120002)

Department of Computer Science and Business System,
Techno International New Town
Kolkata – 700 156
West Bengal, India.

Approval

This is to certify that the project report entitled "Olympics Data Analysis using Azure" prepared under my supervision by Debajit Samanta(18731120017), Subhodeep Roy(18731120002) be accepted in partial fulfillment for the degree of Bachelor of Technology in Computer Science and Business System which is affiliated to Maulana Abul Kalam Azad University of Technology, West Bengal (Formerly known as West Bengal University of Technology).

It is to be understood that by this approval, the undersigned does not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn thereof, but approves the report only for the purpose for which it has been submitted.

…………………………………………
Prof. Bitan Mishra

………………………………………….
Prof. Swagata Paul,
HOD, CSBS/CSE,
Techno International New Town

Abstract

The "Olympics Data Analysis on Microsoft Azure" project is a comprehensive exploration of Olympic datasets leveraging the capabilities of Microsoft Azure's cloud services. The initiative involves raw data ingestion into Azure Data Lake Storage Gen2, subsequent transformation using Azure Databricks for cleansing and filtering, and secure storage back into the Data Lake. Azure Synapse Analytics facilitates advanced analysis with MySQL code, unveiling patterns and insights. Emphasis on documentation and adherence to security and compliance standards ensures transparency and data integrity. The project culminates in visually compelling representations of key findings using Power BI, showcasing proficiency in Azure services for holistic data engineering and analysis.

CONTENTS

LIST OF FIGURES

1. Introduction

In the realm of sports, the Olympic Games stand as a pinnacle of athleticism and international competition, showcasing the world's finest athletes across a diverse range of disciplines. The vast amount of data generated from these competitions holds immense potential for uncovering valuable insights into athlete performance, trends in sports development, and the overall evolution of the Olympic Games. This final year project aims to harness the power of Microsoft Azure, a comprehensive cloud computing platform, to delve into the rich tapestry of Olympic data and extract meaningful knowledge.

The project will commence with the acquisition of Olympic data from a variety of sources, ensuring its preservation in its raw, unfiltered form within Azure Data Lake Gen2. Subsequently, Azure Databricks, a distributed data processing platform, will be employed to perform essential data transformations, cleansing, normalization, and structuring to prepare the data for in-depth analysis. The transformed data will then be meticulously stored back into Azure Data Lake Gen2, maintaining data lineage and ensuring accessibility for downstream applications and analytics tools.

At the heart of the project lies Azure Synapse Analytics, a cloud-based data analytics platform that will empower the exploration of the transformed Olympic data using MySQL code. This comprehensive analysis will delve into identifying trends, patterns, and insights embedded within the data, illuminating aspects such as performance trends over time, medal distribution by country, and demographic trends among athletes. The culmination of this endeavor will manifest in the form of insightful reports, dashboards, and visualizations that effectively communicate the project's findings and contribute to the broader understanding of the Olympic Games.

2. Problem Definition

This project addresses the challenge of efficiently managing and analyzing large-scale Olympic Games data, emphasizing the need for a streamlined and scalable solution. The complexity of raw Olympic data, coupled with the increasing demand for real-time insights, poses a significant hurdle that necessitates advanced cloud-based technologies. The problem revolves around designing an end-to-end data processing pipeline using Microsoft Azure services to ingest, store, and analyze Olympic data, ultimately providing a platform for extracting meaningful patterns and trends. The project aims to overcome the intricacies associated with data volume, diversity, and processing speed, offering a comprehensive solution for researchers, analysts, and enthusiasts to glean valuable insights from the extensive historical and contemporary Olympic datasets.

3. Architecture

This architecture outlines the flow of data from collection to visualization, utilizing Azure Databricks, Data Lake Storage Gen2, and Synapse Analytics in a seamless pipeline.

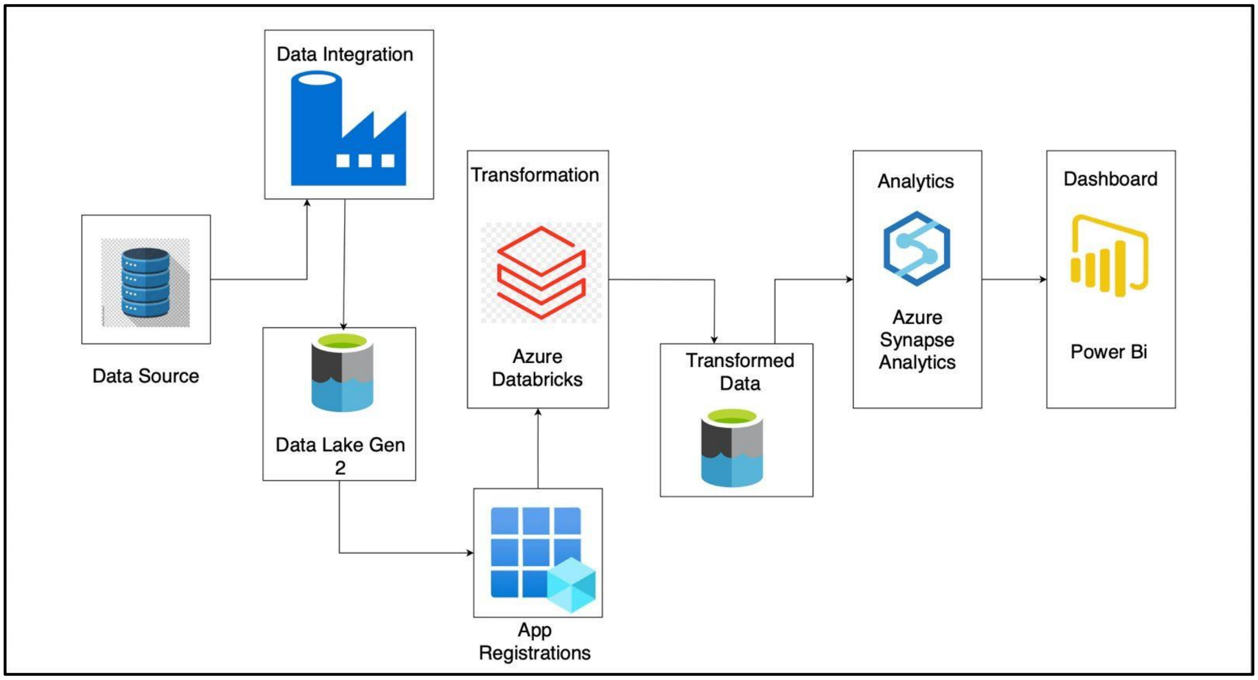


Figure 1. Flowchart of the model of the project.

4. Process and Source Code

This is an optional section for this template. It is applicable if the design has been implemented with some desired output. This section should be framed as follows:

i) **Hardware Requirements**:

a) Desktop with stable internet connection

b) Azure Subscription

ii) Software Requirements:

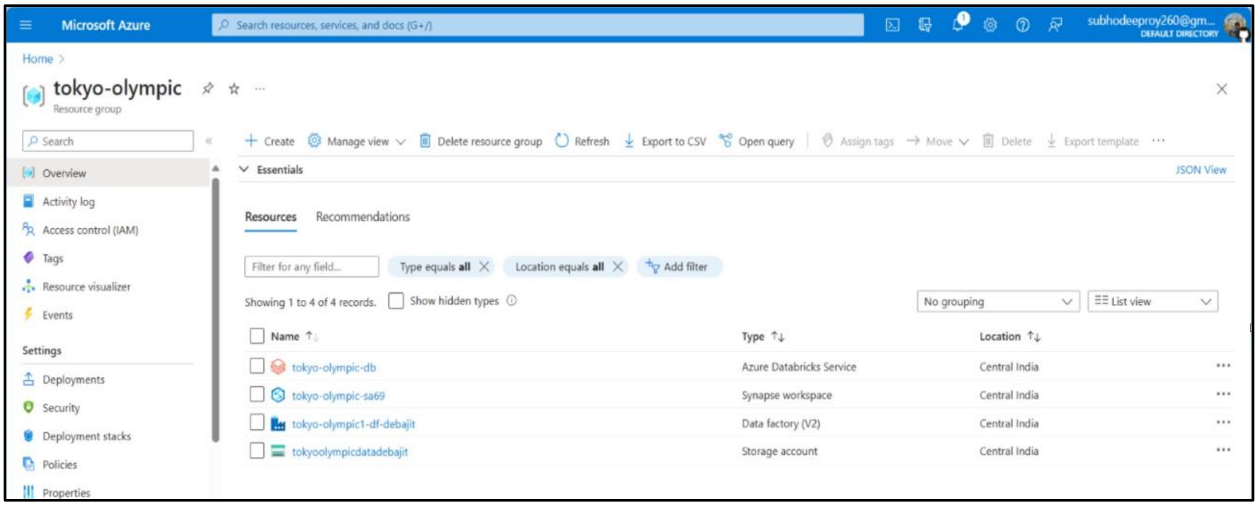
a) Programming Language: Python(Apache Spark), MySQL.

b) Dataset used: Olympics data from Kaggle

iii) Process

a. <u>Data Integration</u>

Create of a data factory



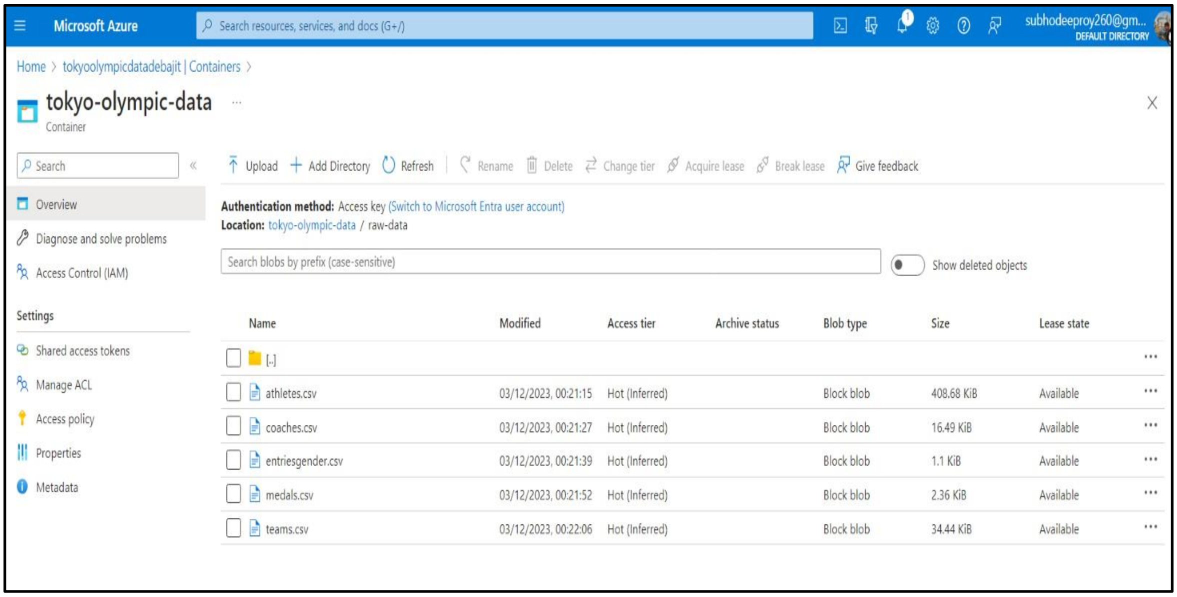
tokyo-olympic
Resource group

Showing 1 to 4 of 4 records. □ Show hidden types

Name	Type	Location
tokyo-olympic-db	Azure Databricks Service	Central India
tokyo-olympic-sa69	Synapse workspace	Central India
tokyo-olympic1-df-debajit	Data factory (V2)	Central India
tokyoolympicdatadebajit	Storage account	Central India

Fig 2. Resource Group

Create a linked service: In the data factory create a linked service to connect to the HTTP data source. Then provide the base URL, authentication type, and other properties of the HTTP data source. Create a linked service to connect to the Azure Gen2 lake storage account where the CSV files will be stored.



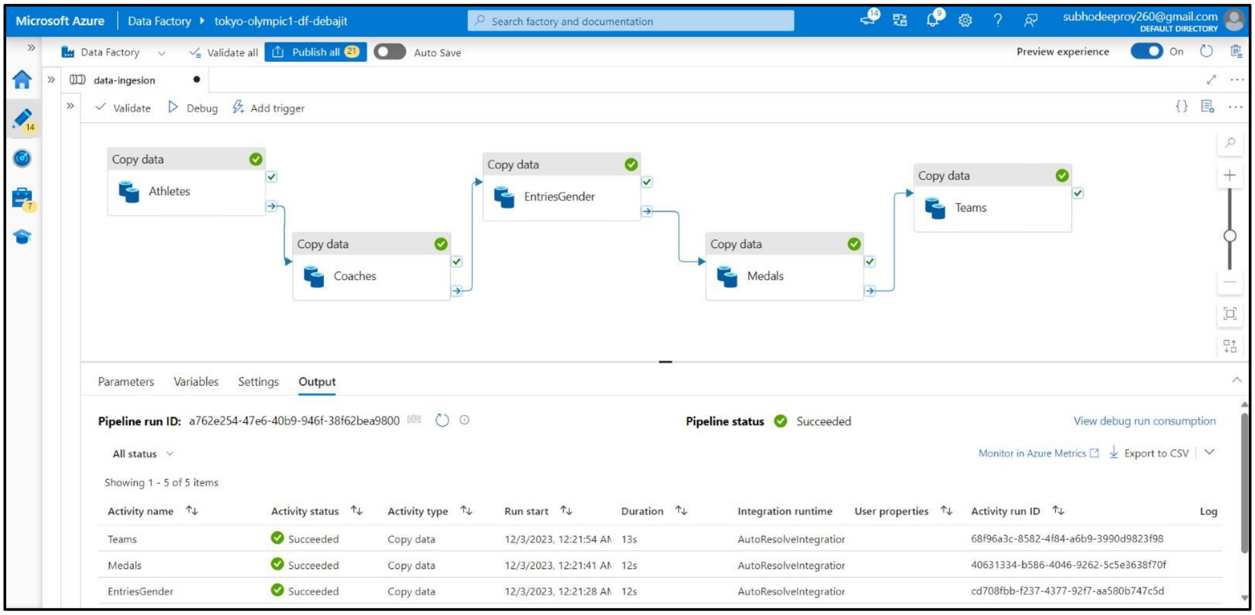
Authentication method: Access key (Switch to Microsoft Entra user account)
Location: tokyo-olympic-data / raw-data

Search blobs by prefix (case-sensitive)

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
[..]						
athletes.csv	03/12/2023, 00:21:15	Hot (Inferred)		Block blob	408.68 KiB	Available
coaches.csv	03/12/2023, 00:21:27	Hot (Inferred)		Block blob	16.49 KiB	Available
entriesgender.csv	03/12/2023, 00:21:39	Hot (Inferred)		Block blob	1.1 KiB	Available
medals.csv	03/12/2023, 00:21:52	Hot (Inferred)		Block blob	2.36 KiB	Available
teams.csv	03/12/2023, 00:22:06	Hot (Inferred)		Block blob	34.44 KiB	Available

Fig 3. Raw data fetched from HTTP and stored.

Create a Pipeline: In the data factory, create a pipeline to orchestrate the data ingestion process. We can use the copy data activity to copy data from the HTTP data source to the Azure Gen2 lake storage account. We will need to configure the source and sink settings, such as the file name, folder path, compression type, and delimiter. We can also use other activities, such as data flow, lookup, or stored procedure, to perform additional transformations or validations on the data.



Pipeline run ID: a762e254-47e6-40b9-946f-38f62bea9800

Pipeline status Succeeded

View debug run consumption

All status

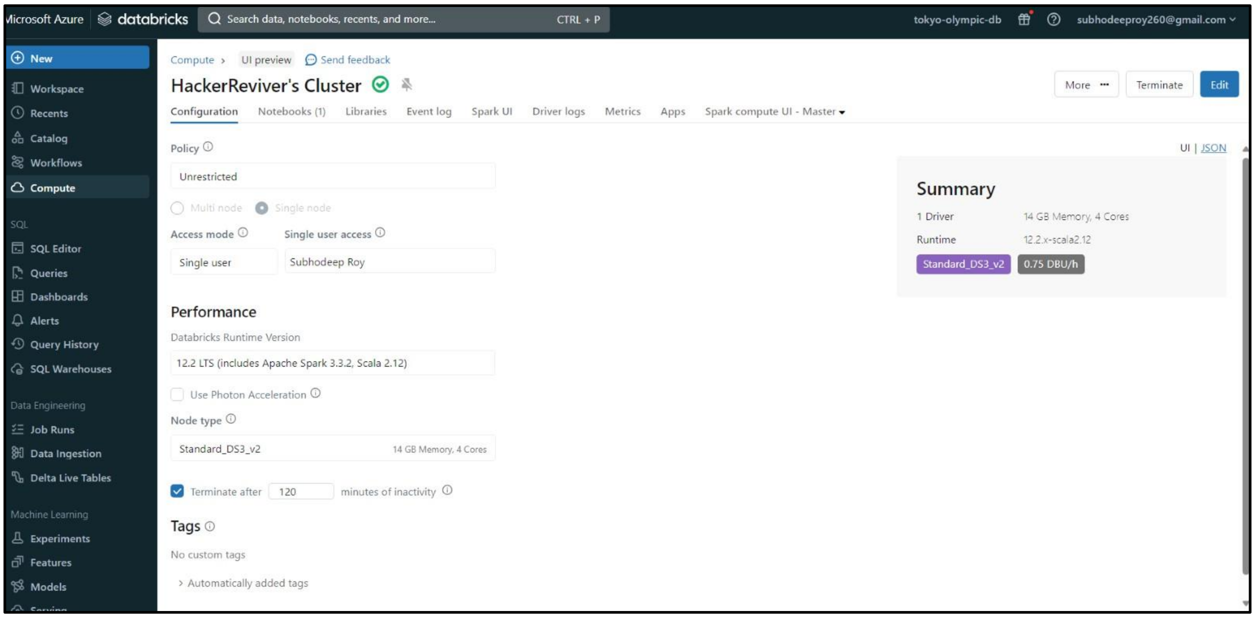
Monitor in Azure Metrics   Export to CSV

Showing 1 - 5 of 5 items

Activity name	Activity status	Activity type	Run start	Duration	Integration runtime	User properties	Activity run ID	Log
Teams	Succeeded	Copy data	12/3/2023, 12:21:54 AN	13s	AutoResolveIntegratior		68f96a3c-8582-4f84-a6b9-3990d9823f98	
Medals	Succeeded	Copy data	12/3/2023, 12:21:41 AN	12s	AutoResolveIntegratior		40631334-b586-4046-9262-5c5e3638f70f	
EntriesGender	Succeeded	Copy data	12/3/2023, 12:21:28 AN	12s	AutoResolveIntegratior		cd708fbb-f237-4377-92f7-aa580b747c5d	

Fig 4. Complete Pipeline

b. Data Processing



HackerReviver's Cluster

Policy: Unrestricted

○ Multi node ● Single node

Access mode: Single user

Single user access: Subhodeep Roy

Performance

Databricks Runtime Version: 12.2 LTS (includes Apache Spark 3.3.2, Scala 2.12)

□ Use Photon Acceleration

Node type: Standard_DS3_v2 14 GB Memory, 4 Cores

☑ Terminate after 120 minutes of inactivity

Tags

No custom tags

› Automatically added tags

Summary

1 Driver 14 GB Memory, 4 Cores

Runtime 12.2.x-scala2.12

Standard_DS3_v2 0.75 DBU/h

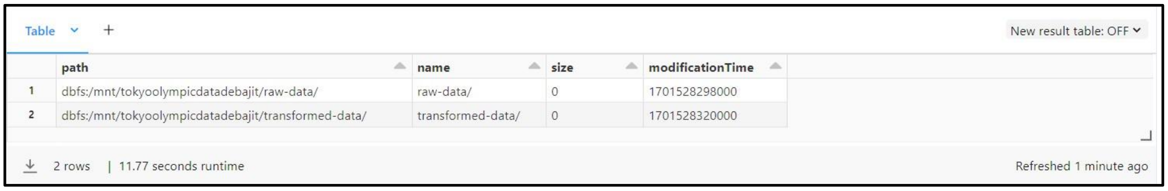
Fig 5. Databricks Cluster

Azure Databricks transformation codes:

```
from pyspark.sql.functions import col

from pyspark.sql.types import IntegerType, DoubleType, BooleanType, DateType



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": 'f0ae821b-15ce-4664-9651-600308006385',

"fs.azure.account.oauth2.client.secret": 'd.g8Q~LezFLBWOSOneBZJXTi3MXqnTd.4yzb0cTI',

"fs.azure.account.oauth2.client.endpoint": "https://login.microsoftonline.com/
a7e93750-926a-46b9-aba4-b8ab061c7fc0/oauth2/token"}
```

```
dbutils.fs.mount(

source = "abfss://tokyo-olympic-data@tokyoolympicdatadebajit.dfs.core.windows.net", # contrainer@storageacc for mount

mount_point = "/mnt/tokyoolympicdatadebajit",

extra_configs = configs)
```

```
%fs

ls "/mnt/tokyoolympicdatadebajit"
```



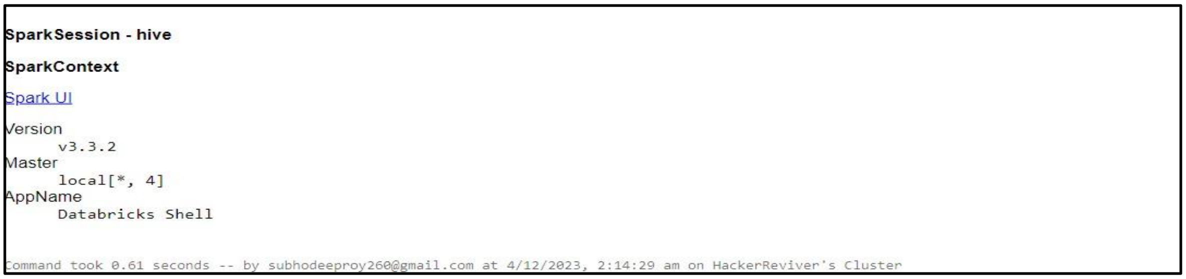
	path	name	size	modificationTime
1	dbfs:/mnt/tokyoolympicdatadebajit/raw-data/	raw-data/	0	1701528298000
2	dbfs:/mnt/tokyoolympicdatadebajit/transformed-data/	transformed-data/	0	1701528320000

2 rows | 11.77 seconds runtime

Refreshed 1 minute ago

Fig 6. Databricks mounted with storage location

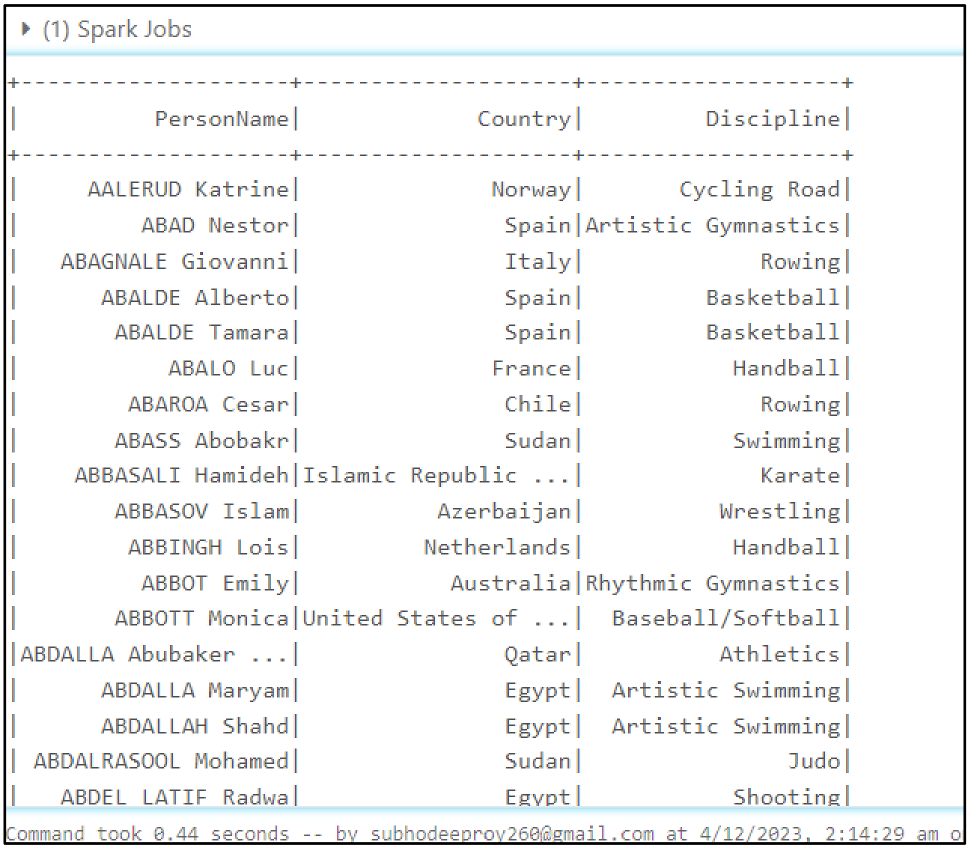
<u>Spark</u>



```
SparkSession - hive

SparkContext

Spark UI

Version
      v3.3.2
Master
      local[*, 4]
AppName
      Databricks Shell


Command took 0.61 seconds -- by subhodeeproy260@gmail.com at 4/12/2023, 2:14:29 am on HackerReviver's Cluster
```

Fig 7. Spark session started

```
athletes = spark.read.format(“csv").option("header","true").option("inferSchema","true").load("/mnt/tokyoolympicdatadebajit/raw-data/athletes.csv")

athelete.show()
```

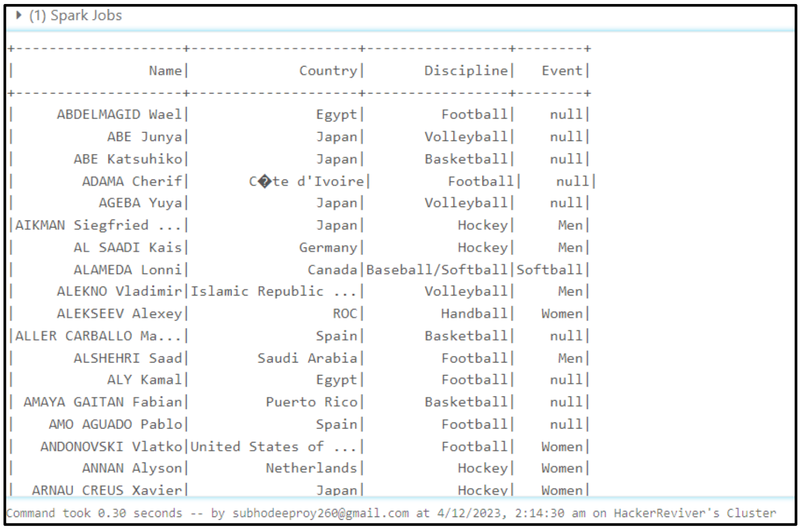


```
+--------------------+--------------------+-------------------+
|          PersonName|             Country|         Discipline|
+--------------------+--------------------+-------------------+
|     AALERUD Katrine|              Norway|       Cycling Road|
|         ABAD Nestor|               Spain|Artistic Gymnastics|
|   ABAGNALE Giovanni|               Italy|             Rowing|
|      ABALDE Alberto|               Spain|         Basketball|
|       ABALDE Tamara|               Spain|         Basketball|
|           ABALO Luc|              France|           Handball|
|        ABAROA Cesar|               Chile|             Rowing|
|        ABASS Abobakr|               Sudan|           Swimming|
|    ABBASALI Hamideh|Islamic Republic ...|             Karate|
|       ABBASOV Islam|          Azerbaijan|          Wrestling|
|        ABBINGH Lois|         Netherlands|           Handball|
|         ABBOT Emily|           Australia|Rhythmic Gymnastics|
|       ABBOTT Monica|United States of ...|  Baseball/Softball|
|ABDALLA Abubaker ...|               Qatar|          Athletics|
|      ABDALLA Maryam|               Egypt|  Artistic Swimming|
|      ABDALLAH Shahd|               Egypt|  Artistic Swimming|
| ABDALRASOOL Mohamed|               Sudan|               Judo|
|   ABDEL LATIF Radwa|               Egypt|           Shooting|
Command took 0.44 seconds -- by subhodeeproy260@gmail.com at 4/12/2023, 2:14:29 am o
```

Fig 8. Atheletes table

coaches=spark.read.format(“csv").option("header","true").option("inferSchema","true").load("/mnt/tokyoolympicdatadebajit/raw-data/coaches.csv")

coaches.show()

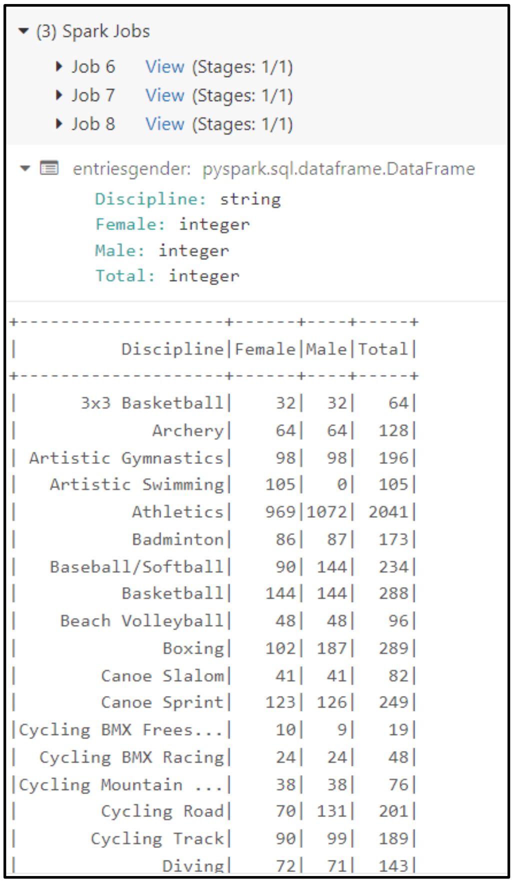


```
+--------------------+--------------------+-----------------+--------+
|                Name|             Country|       Discipline|   Event|
+--------------------+--------------------+-----------------+--------+
|     ABDELMAGID Wael|               Egypt|         Football|    null|
|           ABE Junya|               Japan|       Volleyball|    null|
|       ABE Katsuhiko|               Japan|       Basketball|    null|
|        ADAMA Cherif|       C�te d'Ivoire|         Football|    null|
|          AGEBA Yuya|               Japan|       Volleyball|    null|
|AIKMAN Siegfried ...|               Japan|           Hockey|     Men|
|       AL SAADI Kais|             Germany|           Hockey|     Men|
|       ALAMEDA Lonni|              Canada|Baseball/Softball|Softball|
|     ALEKNO Vladimir|Islamic Republic ...|       Volleyball|     Men|
|     ALEKSEEV Alexey|                 ROC|         Handball|   Women|
|ALLER CARBALLO Ma...|               Spain|       Basketball|    null|
|       ALSHEHRI Saad|        Saudi Arabia|         Football|     Men|
|           ALY Kamal|               Egypt|         Football|    null|
| AMAYA GAITAN Fabian|         Puerto Rico|       Basketball|    null|
|    AMO AGUADO Pablo|               Spain|         Football|    null|
|   ANDONOVSKI Vlatko|United States of ...|         Football|   Women|
|        ANNAN Alyson|         Netherlands|           Hockey|   Women|
|  ARNAU CREUS Xavier|               Japan|           Hockey|   Women|
Command took 0.30 seconds -- by subhodeeproy260@gmail.com at 4/12/2023, 2:14:30 am on HackerReviver's Cluster
```

Fig 9. Coaches table

entriesgender=spark.read.format(“csv").option("header","true").option("inferSchema","true").load("/mnt/tokyoolympicdatadebajit/raw-data/entriesgender.csv")

entriesgender.show()

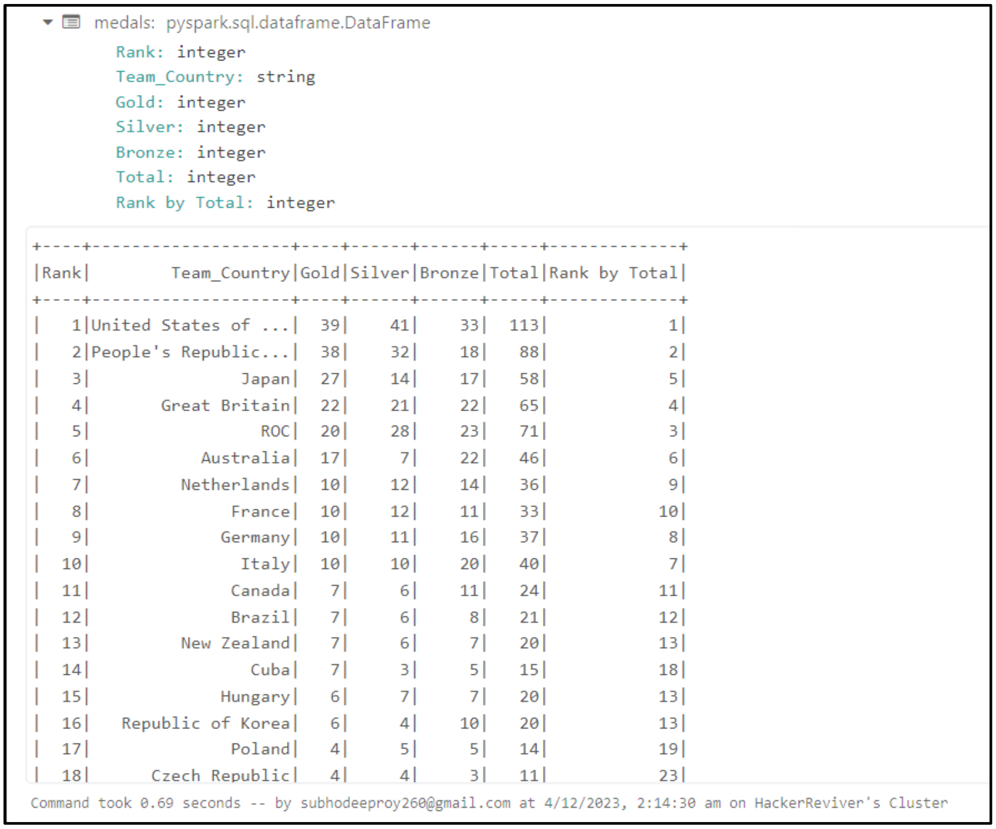


```
▾ (3) Spark Jobs
    ▸ Job 6   View (Stages: 1/1)
    ▸ Job 7   View (Stages: 1/1)
    ▸ Job 8   View (Stages: 1/1)

▾ ▤ entriesgender:  pyspark.sql.dataframe.DataFrame
        Discipline: string
        Female: integer
        Male: integer
        Total: integer

+--------------------+------+----+-----+
|          Discipline|Female|Male|Total|
+--------------------+------+----+-----+
|      3x3 Basketball|    32|  32|   64|
|             Archery|    64|  64|  128|
| Artistic Gymnastics|    98|  98|  196|
|   Artistic Swimming|   105|   0|  105|
|           Athletics|   969|1072| 2041|
|           Badminton|    86|  87|  173|
|   Baseball/Softball|    90| 144|  234|
|          Basketball|   144| 144|  288|
|    Beach Volleyball|    48|  48|   96|
|              Boxing|   102| 187|  289|
|        Canoe Slalom|    41|  41|   82|
|        Canoe Sprint|   123| 126|  249|
|Cycling BMX Frees...|    10|   9|   19|
|  Cycling BMX Racing|    24|  24|   48|
|Cycling Mountain ...|    38|  38|   76|
|        Cycling Road|    70| 131|  201|
|       Cycling Track|    90|  99|  189|
|              Diving|    72|  71|  143|
```

Fig 10. Entriesgender table

medals=spark.read.format("csv").option("header","true").option("inferSchema","true").load("/mnt/tokyoolympicdatadebajit/raw-data/medals.csv")

medals.show()



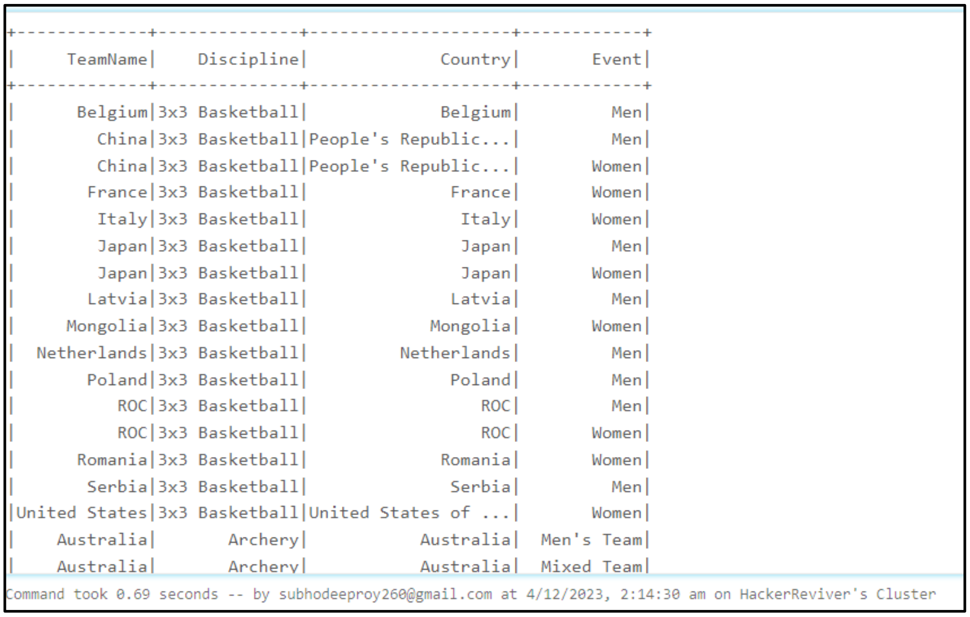
```
medals:  pyspark.sql.dataframe.DataFrame
        Rank:  integer
        Team_Country:  string
        Gold:  integer
        Silver:  integer
        Bronze:  integer
        Total:  integer
        Rank by Total:  integer
```

```
+----+--------------------+----+------+------+-----+-------------+
|Rank|        Team_Country|Gold|Silver|Bronze|Total|Rank by Total|
+----+--------------------+----+------+------+-----+-------------+
|   1|United States of ...|  39|    41|    33|  113|            1|
|   2|People's Republic...|  38|    32|    18|   88|            2|
|   3|               Japan|  27|    14|    17|   58|            5|
|   4|       Great Britain|  22|    21|    22|   65|            4|
|   5|                 ROC|  20|    28|    23|   71|            3|
|   6|           Australia|  17|     7|    22|   46|            6|
|   7|         Netherlands|  10|    12|    14|   36|            9|
|   8|              France|  10|    12|    11|   33|           10|
|   9|             Germany|  10|    11|    16|   37|            8|
|  10|               Italy|  10|    10|    20|   40|            7|
|  11|              Canada|   7|     6|    11|   24|           11|
|  12|              Brazil|   7|     6|     8|   21|           12|
|  13|         New Zealand|   7|     6|     7|   20|           13|
|  14|                Cuba|   7|     3|     5|   15|           18|
|  15|             Hungary|   6|     7|     7|   20|           13|
|  16|   Republic of Korea|   6|     4|    10|   20|           13|
|  17|              Poland|   4|     5|     5|   14|           19|
|  18|      Czech Republic|   4|     4|     3|   11|           23|
Command took 0.69 seconds -- by subhodeeproy260@gmail.com at 4/12/2023, 2:14:30 am on HackerReviver's Cluster
```

Fig11. Medals data

teams=spark.read.format("csv").option("header","true").option("inferSchema","true").load("/mnt/tokyoolympicdatadebajit/raw-data/teams.csv")

teams.show()



```
+-------------+--------------+--------------------+------------+
|     TeamName|    Discipline|             Country|       Event|
+-------------+--------------+--------------------+------------+
|      Belgium|3x3 Basketball|             Belgium|         Men|
|        China|3x3 Basketball|People's Republic...|         Men|
|        China|3x3 Basketball|People's Republic...|       Women|
|       France|3x3 Basketball|              France|       Women|
|        Italy|3x3 Basketball|               Italy|       Women|
|        Japan|3x3 Basketball|               Japan|         Men|
|        Japan|3x3 Basketball|               Japan|       Women|
|       Latvia|3x3 Basketball|              Latvia|         Men|
|     Mongolia|3x3 Basketball|            Mongolia|       Women|
|  Netherlands|3x3 Basketball|         Netherlands|         Men|
|       Poland|3x3 Basketball|              Poland|         Men|
|          ROC|3x3 Basketball|                 ROC|         Men|
|          ROC|3x3 Basketball|                 ROC|       Women|
|      Romania|3x3 Basketball|             Romania|       Women|
|       Serbia|3x3 Basketball|              Serbia|         Men|
|United States|3x3 Basketball|United States of ...|       Women|
|    Australia|       Archery|           Australia|  Men's Team|
|    Australia|       Archery|           Australia|  Mixed Team|
Command took 0.69 seconds -- by subhodeeproy260@gmail.com at 4/12/2023, 2:14:30 am on HackerReviver's Cluster
```

Fig 12. Teams data

athletes.printSchema()

# athletes.show()

coaches.printSchema()

# coaches.show()

entriesgender.printSchema()

# entriesgender.show()

medals.printSchema()

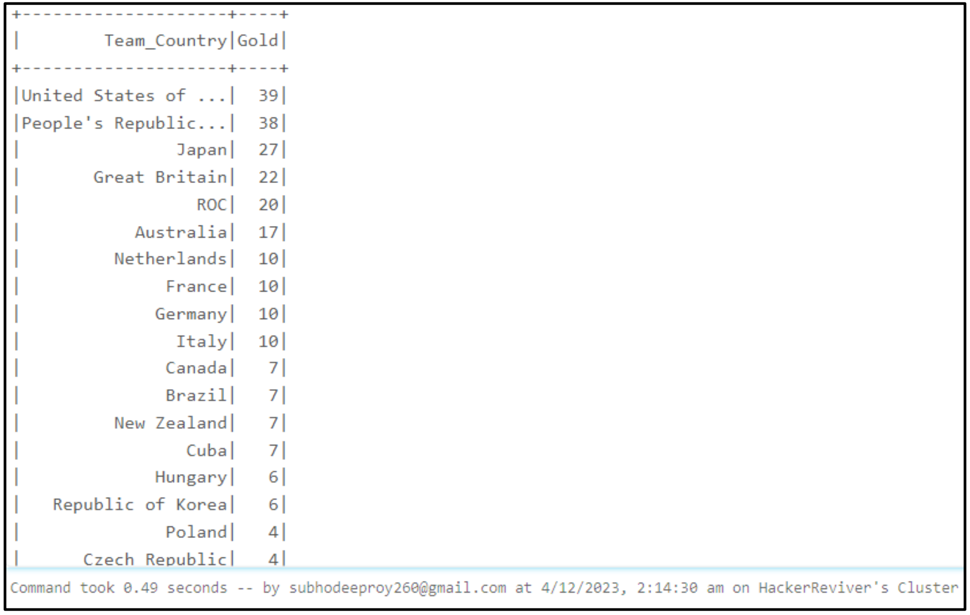
# medals.show()

teams.printSchema()

# teams.show()

### Find the top countries with the highest number of gold medals(using sort function)

top_gold_medal_countries = medals.orderBy("Gold", ascending=False).select(“Team_Country","Gold").show()



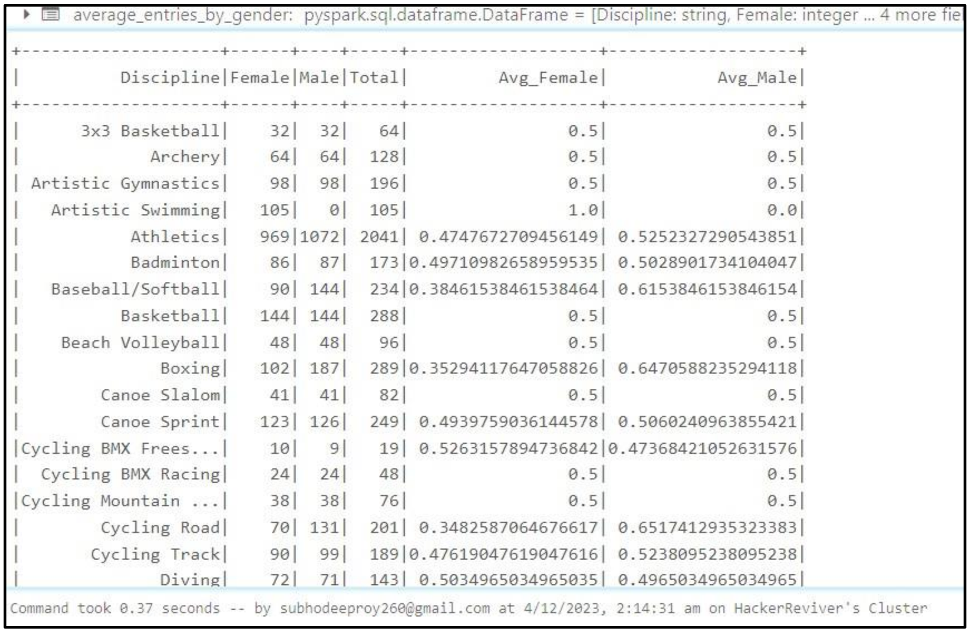
```
+--------------------+----+
|        Team_Country|Gold|
+--------------------+----+
|United States of ...|  39|
|People's Republic...|  38|
|               Japan|  27|
|       Great Britain|  22|
|                 ROC|  20|
|           Australia|  17|
|         Netherlands|  10|
|              France|  10|
|             Germany|  10|
|               Italy|  10|
|              Canada|   7|
|              Brazil|   7|
|         New Zealand|   7|
|                Cuba|   7|
|             Hungary|   6|
|   Republic of Korea|   6|
|              Poland|   4|
|      Czech Republic|   4|
```

Command took 0.49 seconds -- by subhodeeproy260@gmail.com at 4/12/2023, 2:14:30 am on HackerReviver's Cluster

Fig 13. Sorting by medals number

### average no of male and female participate in every match

average_entries_by_gender = entriesgender.withColumn('Avg_Female', entriesgender['Female'] / entriesgender['Total']).withColumn('Avg_Male', entriesgender['Male'] / entriesgender['Total'])
average_entries_by_gender.show()



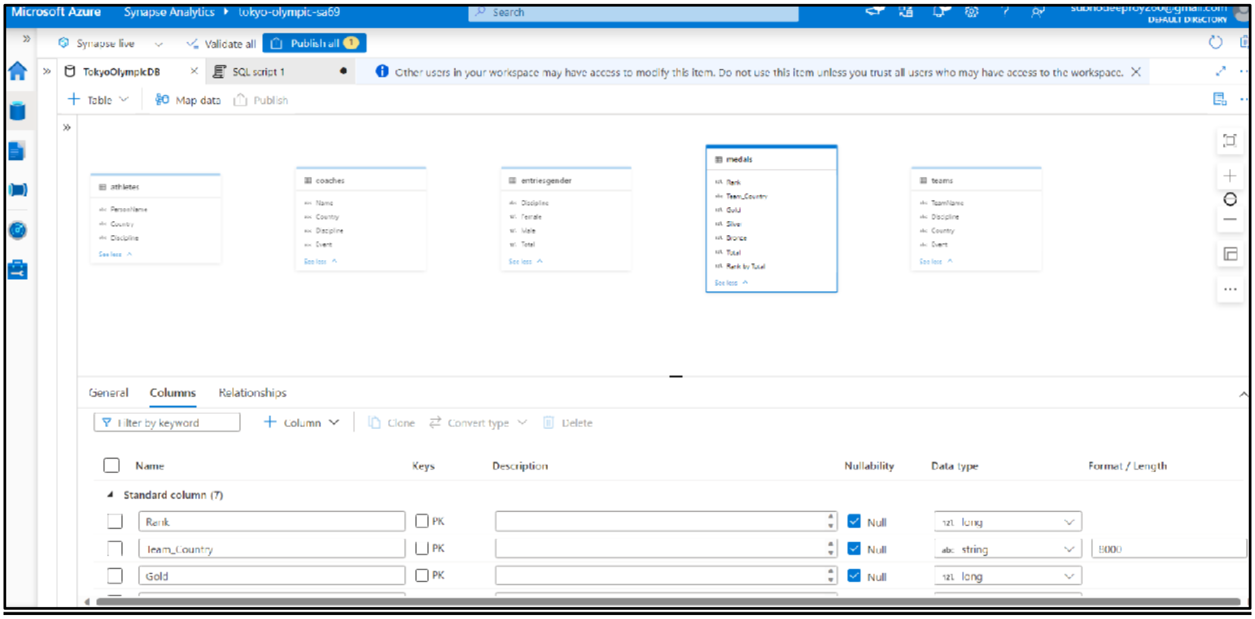
```
average_entries_by_gender:  pyspark.sql.dataframe.DataFrame = [Discipline: string, Female: integer ... 4 more fie

+--------------------+------+----+-----+-------------------+-------------------+
|          Discipline|Female|Male|Total|         Avg_Female|           Avg_Male|
+--------------------+------+----+-----+-------------------+-------------------+
|      3x3 Basketball|    32|  32|   64|                0.5|                0.5|
|             Archery|    64|  64|  128|                0.5|                0.5|
| Artistic Gymnastics|    98|  98|  196|                0.5|                0.5|
|   Artistic Swimming|   105|   0|  105|                1.0|                0.0|
|           Athletics|   969|1072| 2041| 0.4747672709456149| 0.5252327290543851|
|           Badminton|    86|  87|  173|0.49710982658959535| 0.5028901734104047|
|    Baseball/Softball|    90| 144|  234|0.38461538461538464| 0.6153846153846154|
|          Basketball|   144| 144|  288|                0.5|                0.5|
|     Beach Volleyball|    48|  48|   96|                0.5|                0.5|
|              Boxing|   102| 187|  289|0.35294117647058826| 0.6470588235294118|
|        Canoe Slalom|    41|  41|   82|                0.5|                0.5|
|        Canoe Sprint|   123| 126|  249| 0.4939759036144578| 0.5060240963855421|
|Cycling BMX Frees...|    10|   9|   19| 0.5263157894736842|0.47368421052631576|
|   Cycling BMX Racing|    24|  24|   48|                0.5|                0.5|
|Cycling Mountain ...|    38|  38|   76|                0.5|                0.5|
|        Cycling Road|    70| 131|  201| 0.3482587064676617| 0.6517412935323383|
|       Cycling Track|    90|  99|  189|0.47619047619047616| 0.5238095238095238|
|              Diving|    72|  71|  143| 0.5034965034965035| 0.4965034965034965|

Command took 0.37 seconds -- by subhodeeproy260@gmail.com at 4/12/2023, 2:14:31 am on HackerReviver's Cluster
```

Fig 14. Average of female participants

Transformation of datas from raw folder to transformed folder

```
# now transfer the data from source file to raw folder to tranformed folder
athletes.repartition(1).write.mode("overwrite").option("header",'true').csv("/mnt/tokyoolympicdatadebajit/transformed-data/athletes")
coaches.repartition(1).write.mode("overwrite").option("header","true").csv("/mnt/tokyoolympicdatadebajit/transformed-data/coaches")
entriesgender.repartition(1).write.mode("overwrite").option("header","true").csv("/mnt/tokyoolympicdatadebajit/transformed-data/entriesgender")
medals.repartition(1).write.mode("overwrite").option("header","true").csv("/mnt/tokyoolympicdatadebajit/transformed-data/medals")
teams.repartition(1).write.mode("overwrite").option("header","true").csv("/mnt/tokyoolympicdatadebajit/transformed-data/teams")
```

c) Data Analysis



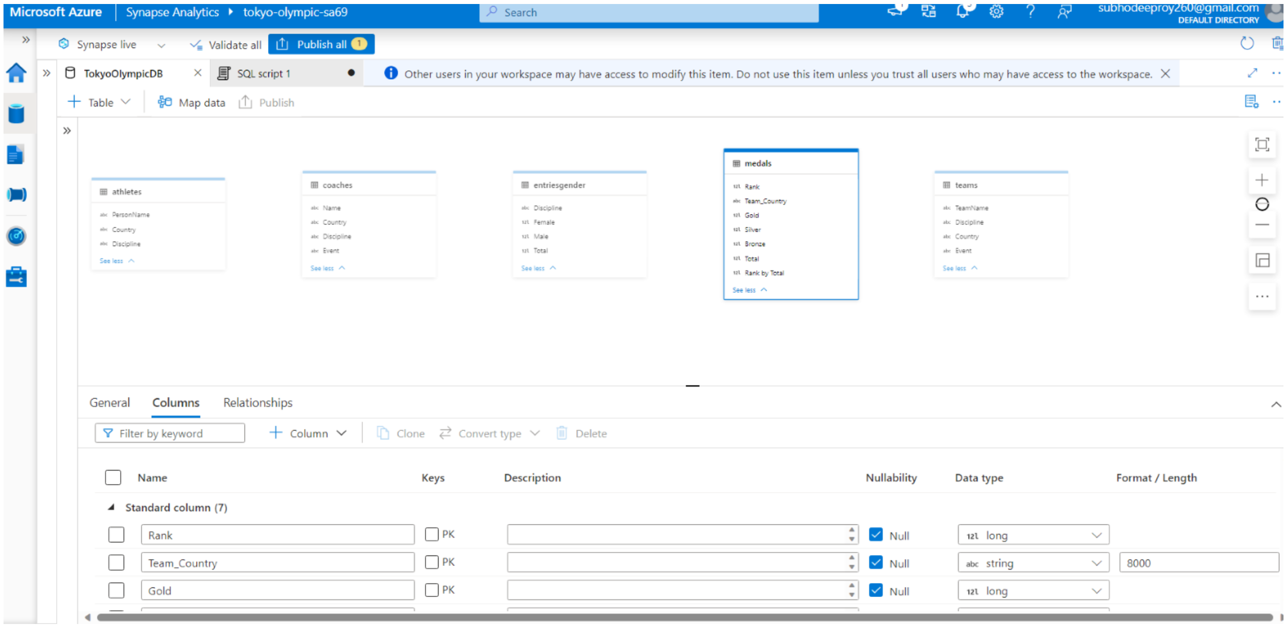
Other users in your workspace may have access to modify this item. Do not use this item unless you trust all users who may have access to the workspace.

	Name	Keys	Description	Nullability	Data type	Format / Length
	Standard column (7)					
□	Rank	□ PK		☑ Null	123 long	
□	Team_Country	□ PK		☑ Null	abc string	8000
□	Gold	□ PK		☑ Null	123 long	

Fig 15. Synapse Analytics

SQL queries in Synapse Analytics

SQL Queries in Synapse Analytics


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.

	Name	Keys	Description	Nullability	Data type	Format / Length
	Standard column (7)					
□	Rank	□ PK		☑ Null	123 long	
□	Team_Country	□ PK		☑ Null	abc string	8000
□	Gold	□ PK		☑ Null	123 long	

```
-- count the no of athletes from each country


 SELECT Country , COUNT(*) AS TotalAthletes

 from athletes
```

```sql
 GROUP BY Country
 ORDER BY TotalAthletes DESC;
```

```sql
--  Calculate the total medals won by each country
SELECT Team_Country,
SUM(Gold) Total_Gold,
SUM(Silver) Total_Silver,
SUM(Bronze)  Total_Bronze
from medals
GROUP BY Team_Country;
```

```sql
---Calculate the average number of entries by gender for each discipline
SELECT Discipline,
AVG (Female) Avg_Female,
AVG(Male) Avg_Male
from entriesgender
GROUP BY Discipline;
```

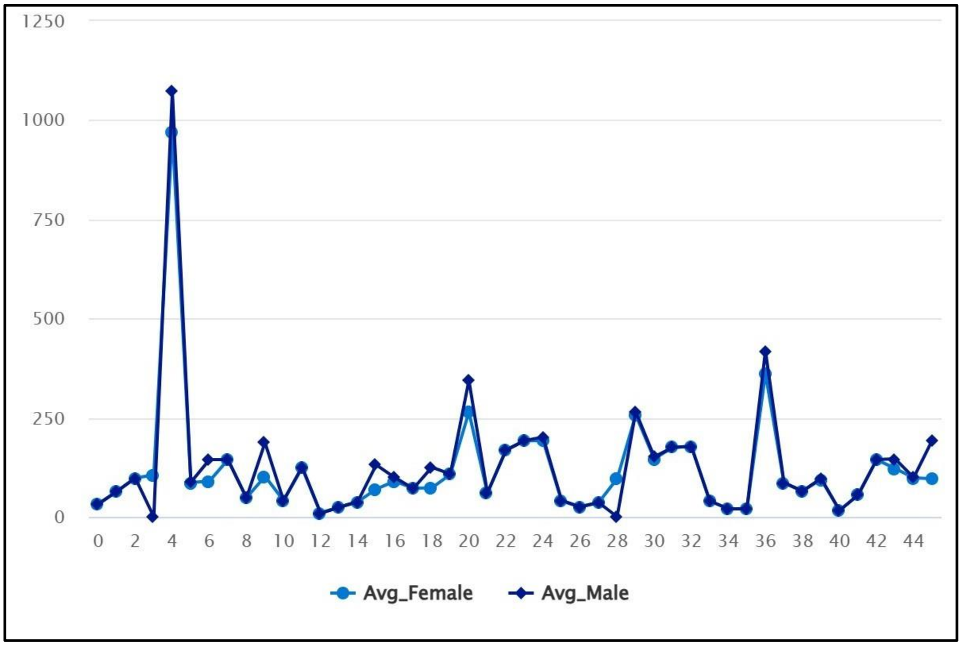


Fig 16. Average number of entries by gender

5. Future Scope of the Project(Should start on a new page)

The future scope of Olympics Data Analysis project can extend in various directions, providing opportunities for ongoing development and exploration. Here are several potential avenues for future enhancements and expansion:

1. Real-Time Data Integration: The incorporation of real-time data streaming to provide up-to-the-minute analytics during ongoing Olympic events. Azure Stream Analytics can be leveraged for this purpose.

2. Machine Learning and Predictive Analytics: Integrate machine learning models to predict future trends, athlete performances, or medal distributions based on historical data. This could involve implementing Azure Machine Learning services.

3. Advanced Visualizations: Enhance the visualization layer with more advanced and interactive dashboards. Using tools like Tableau or advanced Power BI features for richer visual representations.

4. Extended Historical Data Analysis: Expand historical data analysis to cover a more extended period or include additional dimensions, such as geopolitical factors, economic indicators, or social media sentiment during the Olympic events.

5. Geospatial Analysis: Incorporate geospatial analytics to visualize and analyze the geographic distribution of athletes, medal winners, and sporting events. Azure Maps or Power BI's mapping capabilities can be employed for this purpose.

6. Mobile App Integration: Developing a mobile application that provides users with on-the-go access to Olympic data, live updates, and personalized insights. Azure App Service and Xamarin can assist in building cross-platform mobile apps.

7. Social Media Integration: Integrate social media data for sentiment analysis, trending topics, and public reactions related to the Olympic Games. Azure Cognitive Services can be employed for sentiment analysis.

By exploring these avenues, your project can evolve into a dynamic and impactful platform that contributes not only to the understanding of Olympic data but also to the broader field of sports analytics and data science.

6. Conclusion

In conclusion, this project has successfully demonstrated the effective utilization of Microsoft Azure services to address the multifaceted challenges associated with Olympic Games data analysis. From the initial data collection and storage in Azure Databricks and Data Lake Storage Gen2 to the refined analysis using Azure Synapse Analytics, the project has showcased a robust and scalable pipeline for processing large-scale datasets. The application of MySQL code for data analysis and the integration of visualization tools have enabled meaningful insights into various aspects of the Olympic Games. As technology continues to advance, the project's future scope encompasses real-time data integration, machine learning, advanced visualizations, and broader collaborations with sports organizations. The continuous optimization of the data processing pipeline and adherence to data security and privacy measures further contribute to the project's relevance and sustainability. In essence, this endeavor not only signifies a successful implementation of cloud-based data analytics but also opens avenues for ongoing exploration and contribution to the evolving landscape of sports analytics and data science.

7. Bibliography

1. Azure Documentation about Databricks(https://learn.microsoft.com/en-us/azure/databricks/)

2. Azure Documentation about Synapse Analytics(https://learn.microsoft.com/en-us/azure/synapse-analytics/)

3. https://www.youtube.com/

--x---