

# Tokyo Olympics 2020 Data Engineering & Analytics Project

End-to-End Azure Data Engineering Project  
- By Abhishek Nilekar



TOKYO 2020



Microsoft  
Azure



# Index

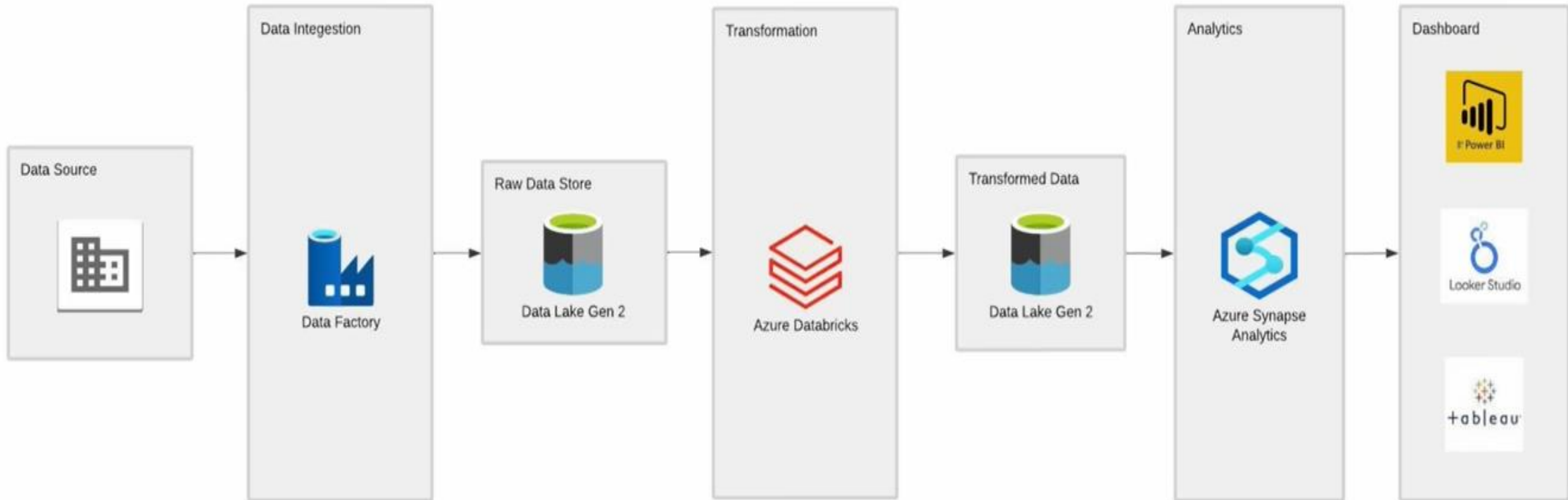
## 1. Project architecture

## 2. Step-wise execution details

- Data Integration
- Data Transformation
- Data Modelling
- Data Analytics
- Visualization and Insights generation

## 3. Key Insights

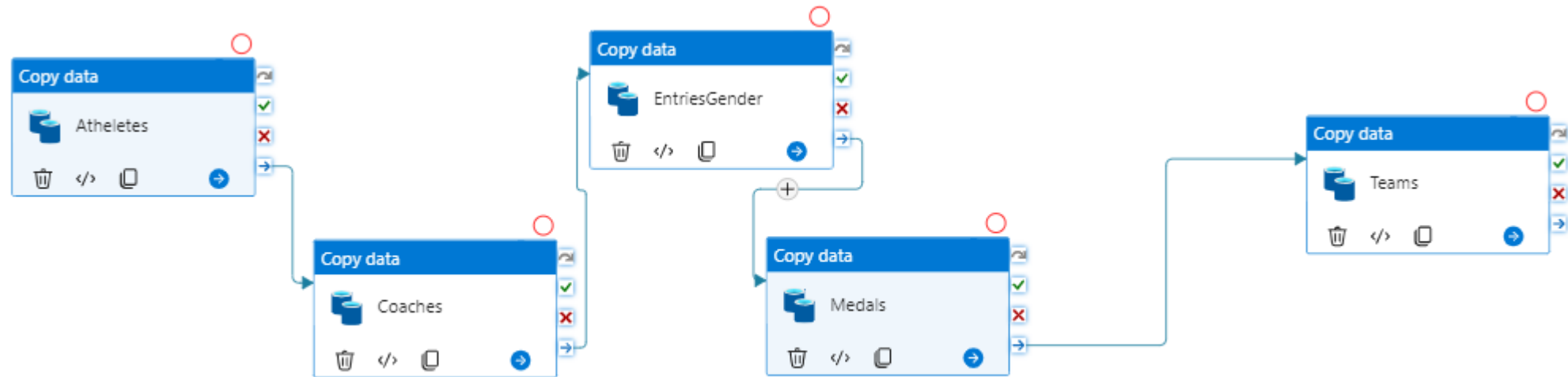
# Project Architecture





# Data Integration using Azure Data Factory

- In the first stage of the project, I integrated data from various sources using Azure Data Factory. This tool allowed me to create data-driven workflows for orchestrating and automating data movement and data transformation.



# Data Transformation using Azure Databricks

- The next step was to transform the data to make it suitable for analysis. I used Azure Databricks for this purpose, which is an Apache Spark-based analytics platform optimized for Azure. It allowed me to transform large volumes of data and make it ready for analysis.

The screenshot shows the Azure Databricks interface for a notebook titled "Tokyo-Olympic-data-transformation-notebook". The interface includes a top bar with "File", "Edit", "View", "Run", and "Help" menus, along with a "Last edit was 22 days ago" timestamp and a "Provide feedback" link. On the right side of the top bar, there are buttons for "Run all", "Terminated", "Schedule", and "Share". The notebook content area displays a command (Cmd 19) with four lines of Python code that use the `repartition` and `write` methods to save transformed data to CSV files. Below the code, it indicates that there are 8 Spark Jobs and that the command took 3.13 seconds to execute. The bottom of the screenshot shows the start of the next command (Cmd 20).

Tokyo-Olympic-data-transformation-notebook Python ☆

File Edit View Run Help Last edit was 22 days ago Provide feedback

► Run all ■ Terminated ▼ 📅 Schedule Share

Cmd 19

```
1 coaches.repartition(1).write.mode("overwrite").option("header","true").csv("/mnt/tokyoolymic/transformed-data/coaches")
2 entriesgender.repartition(1).write.mode("overwrite").option("header","true").csv("/mnt/tokyoolymic/transformed-data/entriesgender")
3 medals.repartition(1).write.mode("overwrite").option("header","true").csv("/mnt/tokyoolymic/transformed-data/medals")
4 teams.repartition(1).write.mode("overwrite").option("header","true").csv("/mnt/tokyoolymic/transformed-data/teams")
```

► (8) Spark Jobs

Command took 3.13 seconds -- by abhisheknilekarde@outlook.com at 11/6/2023, 11:57:56 PM on Abhishek Nilekar's Cluster

Cmd 20



# Data Storage using Azure Data Lake Gen 2

- After integrating the data, I stored it in a raw data store. For this, I used Azure Data Lake Storage Gen2 which provides secure, scalable, and cost-effective storage.
- I stored the transformed data back into Azure Data Lake Storage Gen2. Storing the transformed data separately ensures that the raw data remains untouched and can be used again if needed.

[↑ Upload](#) [Open in Explorer](#) [Delete](#) [→ Move](#) [↺ Refresh](#) [Open in mobile](#) [CLI / PS](#) [Feedback](#)

**i** Enabling SFTP on Azure Blobs has an hourly billing impact. [Learn more about pricing.](#)

^ Essentials

Resource group ([move](#))

: [tokyo-olympic-data](#)

Location

: centralindia

Primary/Secondary Location

: Primary: Central India, Secondary: South India

Subscription ([move](#))

: [Free Trial](#)

Subscription ID

: d3010d30-a79a-4abc-a224-c7e68379d640

Disk state

: Primary: Available, Secondary: Available

Tags ([edit](#))

: [Add tags](#)

Performance

: Standard

Replication

: Read-access geo-redundant storage (RA-GRS)

Account kind

: StorageV2 (general purpose v2)

Provisioning state

: Succeeded

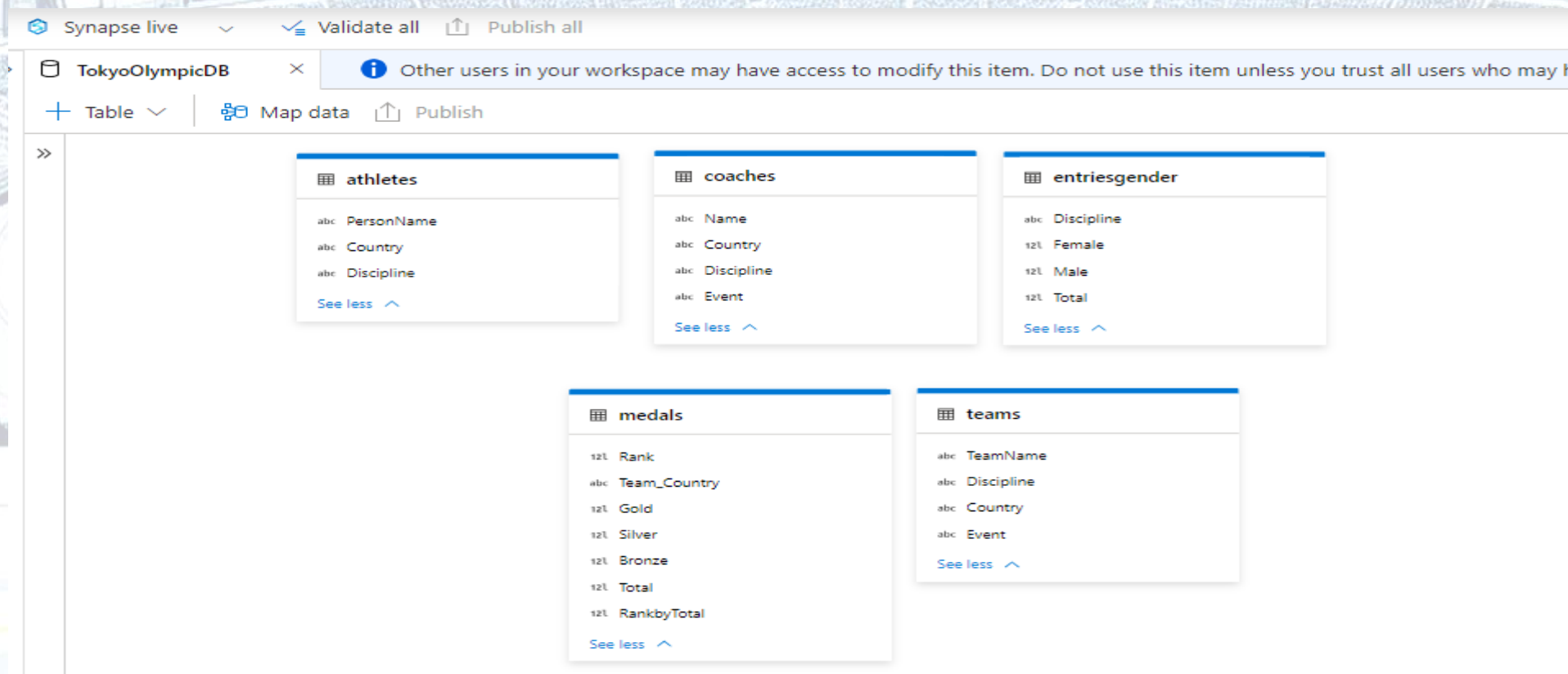
Created

: 11/6/2023, 3:12:35 PM

[JSON View](#)

# Data Analytics using Azure Synapse Analytics

- Finally, I analyzed the data and created a dashboard to visualize it. I used Azure Synapse Analytics, an integrated analytics service, for analyzing the data. For visualization, I used Power BI which helped me create comprehensive dashboards and reports.



The screenshot displays the Azure Synapse Analytics interface. At the top, there's a navigation bar with 'Synapse live', 'Validate all', and 'Publish all' buttons. Below this, a tab for 'TokyoOlympicDB' is active, showing a warning message: '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 it.' The main area shows a data catalog with five tables: athletes, coaches, entriesgender, medals, and teams. Each table is represented by a card with its name and a list of columns.

Table	Columns
athletes	abc PersonName, abc Country, abc Discipline
coaches	abc Name, abc Country, abc Discipline, abc Event
entriesgender	abc Discipline, 121 Female, 121 Male, 121 Total
medals	121 Rank, abc Team_Country, 121 Gold, 121 Silver, 121 Bronze, 121 Total, 121 RankbyTotal
teams	abc TeamName, abc Discipline, abc Country, abc Event





# Data Visualization using Power BI

- For visualization, I used Power BI which helped me create comprehensive dashboards and reports.



## Tokyo Olympics 2020



Total no. of players  
**11.06K**

Total no. of countries  
**206**

Total no. of disciplines  
**46**

### Country-wise medals

Country	Gold medals	Silver medals	Bronze medals	Total medals	Rank by Total medals
United States of America	39	41	33	113	1
People's Republic of China	38	32	18	88	2
ROC	20	28	23	71	3
Great Britain	22	21	22	65	4
Japan	27	14	17	58	5
Australia	17	7	22	46	6
Italy	10	10	20	40	7
Germany	10	11	16	37	8
Netherlands	10	12	14	36	9



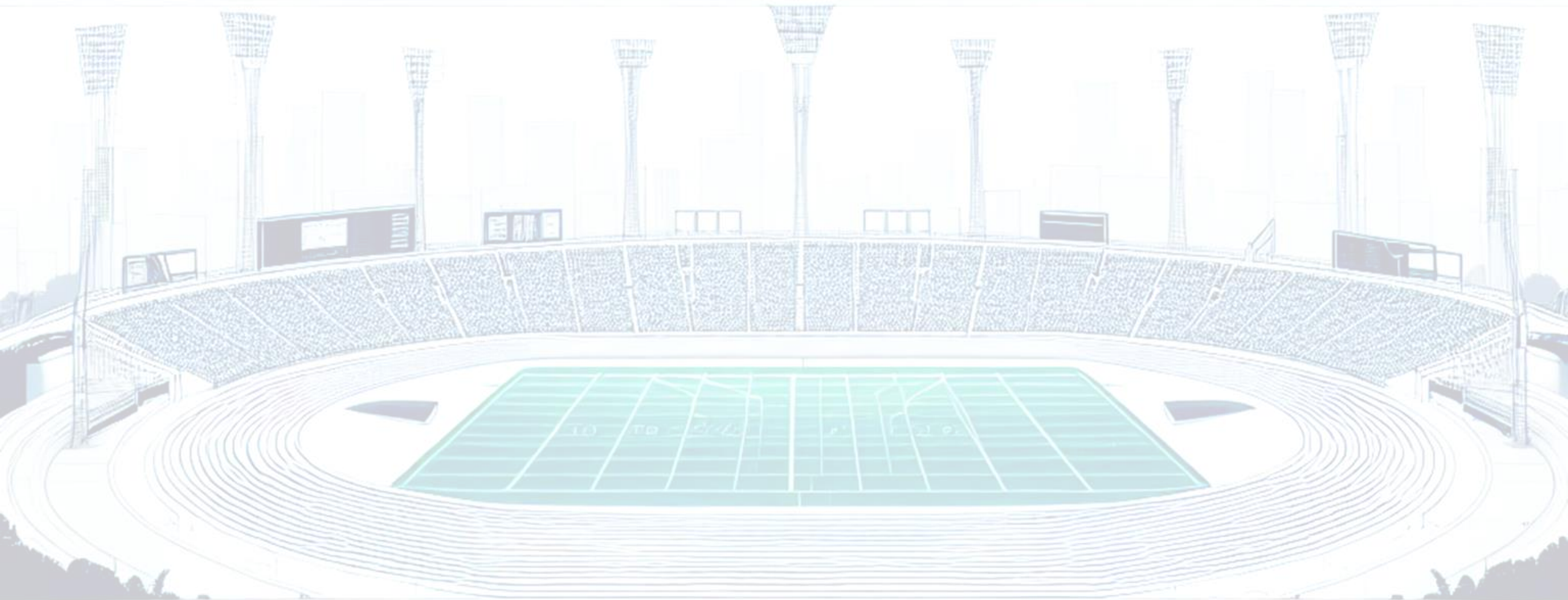
# Key Insights generated from Power BI report

- More than 10 thousand players participated in Tokyo Olympics 2020, the exact number of participants were 11,060 plyers participated from various countries.
- More than 200 countries participated in Tokyo Olympics 2020.
- The total number of disciplines in Tokyo Olympics were 46, some of the important disciplines were:
  - Athletics
  - Swimming
  - Badminton
  - Table Tennis.
- USA topped the medal tally with 113 total medals followed by China with 88 medals followed by Great Britain with 71 total medals.
- India won total 7 medals and ranked 33 among all the countries



# Thank You

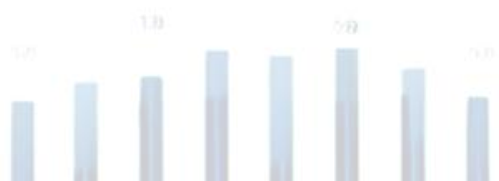




02

03

04



05



06

07



08

