# Olympic Data Analysis Using Azure and Power BI

**Dhanya R Sangolli**

**A20541093**

## Introduction

This report delves into a thorough analysis of Olympic Games data, with a specific focus on assessing athlete performances and the distribution of medals. In this endeavor, we utilize the robust capabilities of Microsoft Azure for data storage, processing, and orchestration, coupled with Power BI for visualization. Our objective is to glean actionable insights from the wealth of historical Olympic records available.

Microsoft Azure: Serving as the cornerstone of our data infrastructure, Azure offers unparalleled scalability and reliability. We leverage Azure Blob Storage as a resilient solution for housing both raw and transformed data, ensuring its accessibility and security. Azure Data Factory plays a pivotal role in streamlining the orchestration of data pipelines, enabling the seamless movement and processing of datasets. Furthermore, Apache Spark on Azure Databricks empowers us to execute large-scale data transformations efficiently, thereby facilitating the extraction of valuable insights from the intricate Olympic datasets.

Medallion Architecture: Our approach adopts the Medallion architecture, comprising three distinct layers: Bronze, Silver, and Gold. Each layer plays a crucial role in ensuring the integrity and usability of the data.

{Bronze Layer}: Serving as the foundational layer, the Bronze layer houses the raw Olympic data, directly ingested into Azure Blob Storage. This layer forms the bedrock upon which subsequent processing relies.

Silver Layer: Positioned as an intermediate stage, the Silver layer is dedicated to cleaning and transforming the data. Here, we meticulously remove duplicate records, fill in missing values, and standardize text fields to ensure consistency. Additionally, we standardize date formats for uniformity across datasets and organize categorical data for enhanced analytical capabilities.

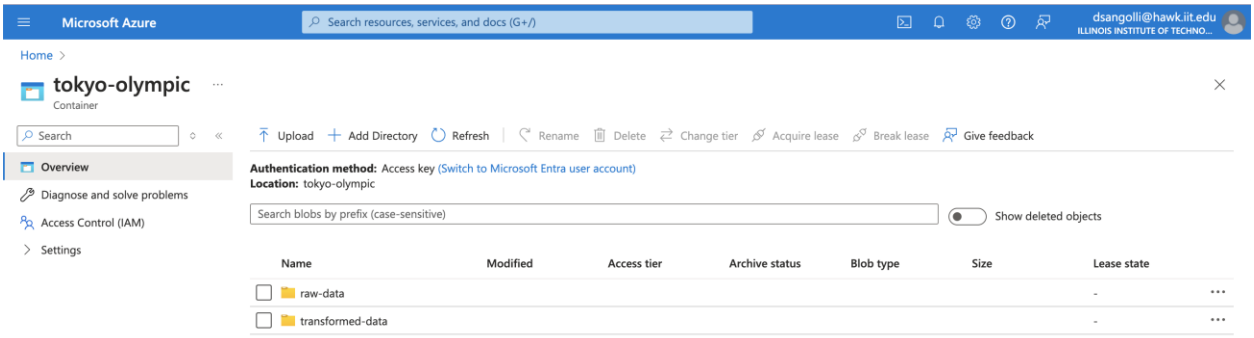
Gold Layer: At the apex of the architecture lies the Gold layer, which comprises structured data in the form of dimension and fact tables. Optimized for in-depth analysis and reporting, dimension tables capture static attributes such as athlete details, coach information, and team profiles, facilitating efficient data slicing and dicing. Conversely, fact tables aggregate measures such as medal counts by country and gender participation across sports, enabling insightful analysis and trend identification.

By harnessing the capabilities of Microsoft Azure and Power BI, this project aims to unlock invaluable insights from Olympic data. The structured Medallion architecture ensures scalability and clarity in data processing and analysis, thereby empowering informed decision-making in the dynamic realm of Olympic sports.

## Tools and Technologies

### Microsoft Azure:

Positioned as the backbone for data storage, processing, and orchestration, Microsoft Azure provides a comprehensive suite of cloud services. Its scalability, reliability, and robustness make it an ideal choice for handling large volumes of data in diverse environments.

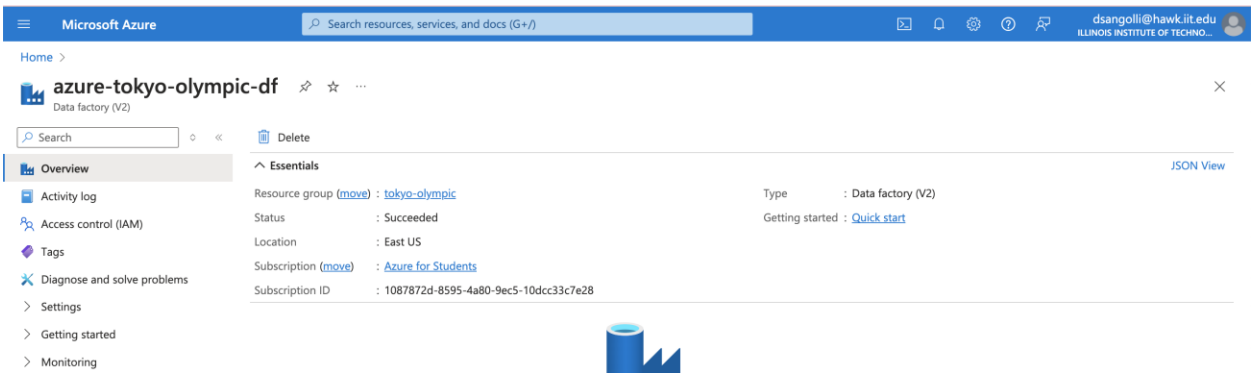



### Azure Blob Storage:

Azure Blob Storage serves as a foundational component, offering a scalable solution for storing both raw and transformed data. Its flexibility and durability make it well-suited for accommodating the diverse needs of data storage in this project.

### Azure Data Factory:

Azure Blob Storage serves as a foundational component, offering a scalable solution for storing both raw and transformed data. Its flexibility and durability make it well-suited for accommodating the diverse needs of data storage in this project.




**Apache Spark on Azure Databricks:**

Apache Spark on Azure Databricks empowers data transformation at scale, enabling the efficient processing of large datasets. By leveraging Spark's distributed computing capabilities within the Azure Databricks environment, complex data transformation tasks can be executed with speed and agility.

**Jupyter Notebook:**

Additionally, Jupyter Notebook has been utilized as a complementary tool for data exploration, analysis, and experimentation. Its interactive and collaborative nature makes it well-suited for prototyping and iterating on data analysis workflows, complementing the capabilities of other tools in the project's toolkit.

**Power BI:**

Power BI is used to enable the creation of interactive dashboards and reports. With its intuitive interface and powerful visualization features, Power BI facilitates the exploration and communication of insights derived from the analyzed Olympic data.

## Data Processing Workflow

### Data Ingestion

Historical Olympic records are ingested into the Bronze layer, providing the raw material for subsequent processing.

## Data Cleaning and Transformation (Silver Layer)

### Cleaning Steps

Duplicate records are removed to ensure data integrity. Missing values are filled, and text fields are standardized for consistency.

### Transformation

Date formats are standardized for uniformity across datasets. Categorical data is organized, and new columns are computed to facilitate analytical queries.

## Data Structuring (Gold Layer)

### Dimension Tables

Static attributes such as athlete details, coach information, and team profiles are extracted and structured into dimension tables. These tables support efficient slicing and dicing of data in reports.

### Fact Tables

Aggregate measures such as medal counts by country and gender participation across sports are computed and stored in fact tables. These tables enable insightful analysis and trend identification.

### Detailed Descriptions and Steps for Dimension and Fact Tables

### Dimension Tables

#### dim_athletes.csv

Attributes:

- PersonName
- AthleteID
- Country
- Discipline

**Description:** Contains detailed information about athletes participating in the Olympics.

- **Extract** athlete names, countries, disciplines, genders, birthdates, heights, weights, and medal counts.
- **Deduplicate** records based on unique athlete identifiers.
- **Standardize** text fields and formats for consistency.

- **Compute** aggregate measures such as total medal counts for each athlete.

**dim_coaches.csv**

Attributes:

- CoachID
- Name
- Country
- Discipline

**Description:** Stores data related to coaches involved in Olympic sports.

- **Capture** coach names, countries, disciplines, genders, and birthdates.
- **Eliminate** duplicate records based on unique coach identifiers.
- **Ensure** consistency in text fields and formats.
- **Optionally**, calculate coaching tenure or other relevant metrics.

**dim_teams.csv**

Attributes:

- TeamID
- TeamName
- Discipline
- Country

**Description:** Provides insights into the teams competing in various sports at the Olympics.

- **Gather** team names, countries, disciplines, and establishment years.
- **Remove** duplicate entries based on unique team identifiers.
- **Standardize** text fields and formats to enhance data consistency.
- **Link** teams to their respective coaches using unique identifiers.

**Fact Tables**

**fact_medal_wins.csv**

Attributes:

- FactID
- CountryID

- DisciplineID
- GoldCount
- BronzeCount
- SilverCount
- TotalMedals

**Description:** Aggregates medal counts by country, reflecting their performance in Olympic events.

- **Group** medal data by country to calculate total gold, silver, and bronze medals.
- **Aggregate** medal counts across disciplines to provide a comprehensive view.
- **Ensure** data accuracy and completeness through rigorous validation.

**fact_entries_gender.csv**

Attributes:

- EntryID
- DisciplineID
- FemaleCount
- MaleCount
- TotalCount

**Description:** Summarizes gender participation across different sports at the Olympics.

- **Tabulate** the number of male and female participants in each sport discipline.
- **Aggregate** gender participation data to analyze overall trends.
- **Validate** data integrity to ensure accurate representation.

**Visualization and Reporting in Power BI**

**Dashboards Created**

**1. Bar Graph: Medal Counts by** Country

- **Description**: This bar graph displays the sum of bronze, silver, and gold medals won by each country. It allows users to quickly assess which countries have the highest medal counts.
- **Data Used**: **fact_medal_wins.csv**
- **Fields**:

    - X-axis: **Country_Name**
    - Y-axis: **BronzeCount**, **SilverCount**, **GoldCount**
- **Visualization Type**: Bar Graph

**2. Stacked Bar Graph: Coach Count by Country**

- **Description**: This stacked bar graph represents the count of coaches by country, segmented by sport discipline. It provides insights into the distribution of coaching resources across different sports within each country.

**3. Pie Chart: Participant Count by Team**

- **Description**: This pie chart shows the count of participants by team name. It is useful for identifying the teams with the highest number of participants, highlighting prominent teams in the Olympics.

**4. Table: Total Medals by Country**

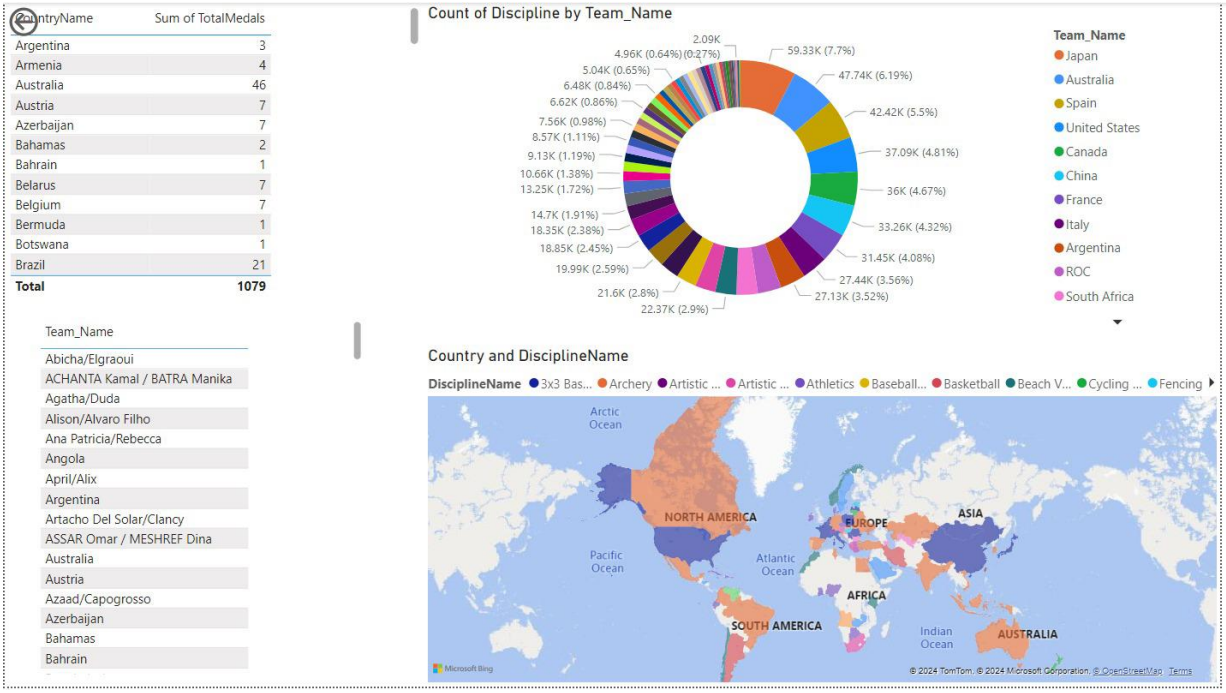
- **Description**: This table displays a summary of total medals by country name, providing a detailed breakdown of medal types (gold, silver, bronze).

**5. Donut Chart: Discipline Count by Team**

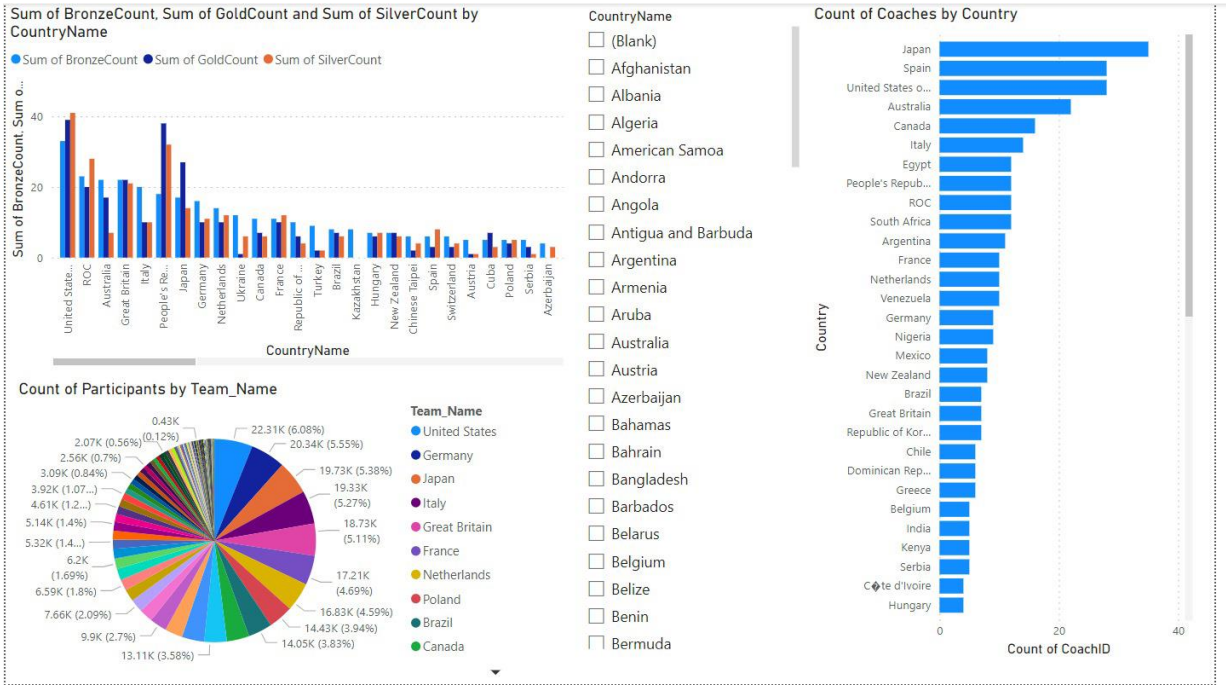
- **Description**: The donut chart visualizes the count of different disciplines represented by each team. It offers a clear view of the diversity of sports that teams are participating in.

**6. Filled Map: Discipline and Country Distribution**

- **Description**: This filled map provides a geographical visualization of countries and the disciplines they participate in, colored by the number of participants in each discipline. It helps visualize the global distribution of sports participation at the Olympic level.
- Dashboard 1


CountryName
Sum of TotalMedals
Argentina 3
Armenia 4
Australia 46
Austria 7
Azerbaijan 7
Bahamas 2
Bahrain 1
Belarus 7
Belgium 7
Bermuda 1
Botswana 1
Brazil 21
Total 1079
Team_Name
Abicha/Elgraoui
ACHANTA Kamal / BATRA Manika
Agatha/Duda
Alison/Alvaro Filho
Ana Patricia/Rebecca
Angola
April/Alix
Argentina
Artacho Del Solar/Clancy
ASSAR Omar / MESHREF Dina
Australia
Austria
Azaad/Capogrosso
Azerbaijan
Bahamas
Bahrain
Count of Discipline by Team_Name
2.09K
4.96K (0.64%)
5.04K (0.65%)
6.48K (0.84%)
6.62K (0.86%)
7.56K (0.98%)
8.57K (1.11%)
9.13K (1.19%)
10.66K (1.38%)
13.25K (1.72%)
14.7K (1.91%)
18.35K (2.38%)
18.85K (2.45%)
19.99K (2.59%)
21.6K (2.8%)
22.37K (2.9%)
59.33K (7.7%)
47.74K (6.19%)
42.42K (5.5%)
37.09K (4.81%)
36K (4.67%)
33.26K (4.32%)
31.45K (4.08%)
27.44K (3.56%)
27.13K (3.52%)
Team_Name
Japan
Australia
Spain
United States
Canada
China
France
Italy
Argentina
ROC
South Africa
Country and DisciplineName
DisciplineName
3x3 Bas...
Archery
Artistic ...
Artistic ...
Athletics
Baseball...
Basketball
Beach V...
Cycling ...
Fencing
Arctic Ocean
NORTH AMERICA
EUROPE
ASIA
Pacific Ocean
Atlantic Ocean
AFRICA
SOUTH AMERICA
Indian Ocean
AUSTRALIA
Microsoft Bing
© 2024 TomTom, © 2024 Microsoft Corporation, © OpenStreetMap Terms

## Dashboard 2


Sum of BronzeCount, Sum of GoldCount and Sum of SilverCount by CountryName
Sum of BronzeCount
Sum of GoldCount
Sum of SilverCount
Sum of BronzeCount, Sum o...
CountryName
United State...
ROC
Australia
Great Britain
Italy
People's Re...
Japan
Germany
Netherlands
Ukraine
Canada
France
Republic of ...
Turkey
Brazil
Kazakhstan
Hungary
New Zealand
Chinese Taipei
Spain
Switzerland
Austria
Cuba
Poland
Serbia
Azerbaijan
Count of Participants by Team_Name
0.43K (0.12%)
2.07K (0.56%)
2.56K (0.7%)
3.09K (0.84%)
3.92K (1.07...)
4.61K (1.2...)
5.14K (1.4%)
5.32K (1.4...)
6.2K (1.69%)
6.59K (1.8%)
7.66K (2.09%)
9.9K (2.7%)
13.11K (3.58%)
22.31K (6.08%)
20.34K (5.55%)
19.73K (5.38%)
19.33K (5.27%)
18.73K (5.11%)
17.21K (4.69%)
16.83K (4.59%)
14.43K (3.94%)
14.05K (3.83%)
Team_Name
United States
Germany
Japan
Italy
Great Britain
France
Netherlands
Poland
Brazil
Canada
CountryName
(Blank)
Afghanistan
Albania
Algeria
American Samoa
Andorra
Angola
Antigua and Barbuda
Argentina
Armenia
Aruba
Australia
Austria
Azerbaijan
Bahamas
Bahrain
Bangladesh
Barbados
Belarus
Belgium
Belize
Benin
Bermuda
Count of Coaches by Country
Japan
Spain
United States o...
Australia
Canada
Italy
Egypt
People's Repub...
ROC
South Africa
Argentina
France
Netherlands
Venezuela
Germany
Nigeria
Mexico
New Zealand
Brazil
Great Britain
Republic of Kor...
Chile
Dominican Rep...
Greece
Belgium
India
Kenya
Serbia
C♦te d'Ivoire
Hungary
Country
Count of CoachID

## Conclusion

By leveraging Azure and Power BI, this project successfully unlocks valuable insights from Olympic data, demonstrating the effectiveness of cloud and BI technologies in handling large datasets. The structured Medallion architecture ensures scalability and clarity in data processing and analysis, facilitating informed decision-making in the realm of Olympic sports.