Olympic Data Analysis Using Azure and Power BI

[Dhanya R Sangolli]

February 4, 2025

Contents

Chapter 1

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.

Chapter 2

Tools and Technologies

2.1 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.

2.2 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.

2.3 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.

2.4 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.

2.5 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.

2.6 Power BI

Power BI is utilized for its capability 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.

Chapter 3

Data Processing Workflow

3.1 Data Ingestion

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

3.2 Data Cleaning and Transformation (Silver Layer)

3.2.1 Cleaning Steps

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

3.2.2 Transformation

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

3.3 Data Structuring (Gold Layer)

3.3.1 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.

3.3.2 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.

Chapter 4

Detailed Descriptions and Steps for Dimension and Fact Tables

4.1 Dimension Tables

4.1.1 dim_athletes.csv

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.

4.1.2 dim_coaches.csv

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.

4.1.3 dim_teams.csv

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.

4.2 Fact Tables

4.2.1 fact_medal_wins.csv

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.

4.2.2 fact_entries_gender.csv

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.

Dimension Tables

dim_athletes.csv

- **athlete_name**: Name of the athlete.
- **Country**: Country represented by the athlete.
- **Discipline**: Sport discipline in which the athlete competes.

dim_coaches.csv

- **coach_name**: Name of the coach.
- **Country**: Country associated with the coach.
- **Discipline**: Sport discipline in which the coach is involved.

dim_teams.csv

- **team_name**: Name of the team.
- **Discipline**: Sport discipline in which the team participates.
- **Country**: Country represented by the team.

Fact Tables

fact_medal_wins.csv

- **Team_Country**: Country of the team.
- **total_gold**: Total number of gold medals won by the team.
- **total_silver**: Total number of silver medals won by the team.
- **total_bronze**: Total number of bronze medals won by the team.

fact_entries_gender.csv

- **Discipline**: Sport discipline.
- **total_male_entries**: Total number of male participants in the discipline.
- **total_female_entries**: Total number of female participants in the discipline.

Chapter 5

Visualization and Reporting in Power BI

5.1 Dashboards Created

5.1.1 Medal Dashboard

Visualizes total medal counts and types by country.

5.1.2 Athlete Performance

5.1.3 Participation by Gender

Shows the distribution of participants by gender in each discipline.

Chapter 6

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.