# OLYMPIC GAMES ANALYTICS USING APACHE SPARK DATABRICKS

Submitted in partial fulfillment of the requirements of the degree

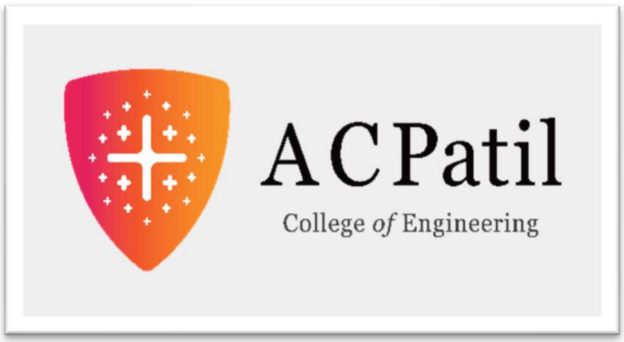
**BACHELOR OF ENGINEERING IN ARTIFICIAL INTELLIGENCE AND DATA SCIENCE**

By


**1. Sachin Bade (07 / 211101003)**
**2. Aarohi Pisolkar (46 / 211101009)**
**3. Arya Sonawane (53 / 221102004)**
**4. Yash Shirsath (65 / 201101006)**


**Project Guide**

**Prof. Veena Bhamre**





**Department of Artificial Intelligence & Data Science**

**A. C. Patil College of Engineering**
**Kharghar, Navi Mumbai**


**University of Mumbai**
**(AY 2024-25)**

# CERTIFICATE

This is to certify that the Mini Project entitled

**OLYMPIC GAMES ANALYTICS USING APACHE SPARK DATABRICKS**

is a Bonafide Work of

**Sachin Bade - 02**
**Aarohi Pisolkar - 46**
**Arya Sonawane - 53**
**Yash Shirsath - 65**

Submitted to the **University of Mumbai** in partial fulfillment of the requirement for the award of the Degree of **"Bachelor of Engineering"** in **"Artificial Engineering and Data Science".**

**Prof. Veena Bhamre**
**(Project Guide)**

**Prof. Shilpali P. Bansu**
(**Head of Department)**

**Dr. V. N. Pawar**
(**Principal)**

# Mini Project Approval

This Mini Project Entitled **"**Olympic Games Analytics Using Apache Spark Databricks**" by Sachin Bade** (2), **Aarohi Pisolkar** (46), **Arya Sonawane** (53), **Yash Shirsath** (65) is Approved for the Degree of Bachelor of Engineering in Artificial Intelligence and Data Science.

**Examiners**

**1……**
(Internal Examiner Name & Sign)

**2……**
(External Examiner name & Sign)

**Date:** 17/10/2024

**Place**: A. C. Patil College of Engineering, Kharghar

# Contents

**Abstract**

**Acknowledgements**

**List of Figures**

**1] Introduction** **9**

**1.1** Introduction

**1.2** Motivation

**1.3** Problem Statement & Objectives

**1.4** Organization of the Report

**2] Literature Survey** **11**

**1.5** Survey of Existing System

**1.6** Limitation Existing system or research gap

**1.7** Mini Project Contribution

**3] Proposed System** **12**

3.1. Introduction
3.2. Architecture/ Framework
3.3. Algorithm and Process Design
3.4. Experiment and Results
3.5. Conclusion and Future work.
3.6. References **19**

# Abstract

The analysis of Olympic Games data offers a fascinating insight into over a century of global sports performance, country participation, and evolving athlete demographics. This project focuses on the comprehensive analysis of Olympic Games data from 1896 to 2016 using Apache Spark and Databricks. The primary objective of this study is to explore trends in athlete performance, country participation, and gender representation, while also utilizing the capabilities of big data tools for efficient data processing and visualization. using Apache Spark, we handle large datasets, perform complex transformations, and derive insightful summaries of the data, such as the top-performing countries by medal counts, participation trends across years, and the impact of gender diversity on sports participation. Databricks, a powerful cloud platform, is employed for real-time visualization of the processed data, helping to identify significant patterns in athlete performance, gender participation, and changes in event types over time.

The project's results highlight the dominance of certain countries in specific sports, reveal shifts in athlete physical attributes (height, weight) over the years, and track the progressive increase in female athlete participation. These insights not only emphasize the utility of modern big data tools in sports analytics but also present opportunities for further research into athlete performance optimization and historical trend analysis in global sporting events.

# Acknowledgment

We would like to express our deepest gratitude to our respected Principal, Dr. V. N. Pawar, whose continued support and provision of resources have made this project possible. We extend special thanks to our Head of the Department, Shilpali Bansu, for her insightful recommendation of this project topic, which holds great significance for the department's academic endeavors. a heartfelt thank you goes to our Project Guide, Prof. Veena Bhamre, for her invaluable mentorship, expert advice, and relentless encouragement throughout the development of this project. Her innovative ideas and guidance have been crucial in shaping this project's success. We are also grateful to all the faculty members for their continuous support, academic insights, and motivation that have contributed greatly to the project's outcome. In addition, we would like to acknowledge the unconditional love and encouragement of our families and friends, who have been our constant source of strength and support. Their understanding and belief in us have been instrumental in the successful completion of this project. this acknowledgment is a small token of our appreciation for all the support, guidance, and encouragement we have received from everyone who has been a part of this journey.

Date: - /10/2024

Sign: -

# List of Figures


| Sr. No. | Name of Picture | Page no. |
|---|---|---|
| 1 | Architecture Diagram for Olympic Games Analytics | 12 |
| 2 | Gold Medals for Athletes Over 50 based on Sports | 12 |
| 3 | Women medals per edition (Summer Season) of the Games | 12 |
| 4 | Top 5 Gold Medal Countries | 12 |
| 5 | Height vs Weight of Olympic Medalists | 13 |
| 6 | Variation of Male Athletes over Time | 13 |

# Chapter 1 – Introduction

## 1.1. Introduction

The Olympic Games, the most prestigious international sporting event, have been held every four years since 1896. They bring together athletes from around the world to compete in a wide range of sports, showcasing athleticism, diversity, and global unity. The rich history and sheer volume of data generated from these events present a unique opportunity for analysis, revealing fascinating insights into the evolution of sports, country participation, and athlete performance over time. this project focuses on the analysis of Olympic Games data using Apache Spark and Databricks, powerful tools designed to handle large datasets and perform complex data operations. By utilizing these technologies, we aim to explore key trends and patterns in Olympic Games history, such as medal distribution, gender participation, and changes in athlete demographics.

The project also emphasizes the importance of big data analytics in deriving meaningful insights from vast datasets. With the help of Apache Spark's distributed computing capabilities and Databricks' interactive environment, we can process and visualize Olympic Games data effectively, demonstrating the real-world application of big data in sports analytics. This introduction lays the foundation for understanding how modern data analysis techniques can unlock valuable information hidden within historical Olympic data.

## 1.2. Motivation

The motivation behind this project stems from the increasing significance of data analytics in various fields, particularly in sports. The Olympic Games represent a unique confluence of history, culture, and athletic achievement, and the wealth of data generated during these events presents an unparalleled opportunity for exploration and analysis. understanding the dynamics of the Olympic Games can reveal insights into athlete performance, national dominance in various sports, and the changing landscape of gender representation in athletics. With the growing emphasis on data-driven decision-making in sports management, coaching, and athlete training, this project aims to harness the power of big data analytics to uncover patterns and trends that could influence future strategies and policies in sports.

Moreover, the project is motivated by the desire to demonstrate the practical applications of modern technologies, such as Apache Spark and Databricks, in managing and analyzing large datasets. By leveraging these tools, we can efficiently process and visualize complex data, showcasing the potential of big data analytics in generating actionable insights. ultimately, this project seeks to contribute to the field of sports analytics by providing a comprehensive analysis of Olympic Games data, fostering a deeper understanding of the factors that shape athletic performance and participation over time. Through this exploration, we hope to inspire further research and discussions around the intersection of sports, data, and technology.

## 1.3. Problem Statement and Objective

- **Problem Statement**: -

The Olympic Games have a rich history and generate vast amounts of data over each cycle, including athlete performance, country participation, medal distribution, and demographic information. However, extracting meaningful insights from this extensive dataset can be challenging due to its complexity and size. Traditional data analysis methods often fall short when handling such large volumes of information, leading to missed opportunities for understanding key trends and patterns that could inform future sporting events, athlete training, and national sports policies. the primary challenge lies in efficiently processing and analysing this large dataset to reveal valuable insights that reflect the changing dynamics of the Olympic Games over time. Additionally, there is a need for better visualization techniques to present these insights in a comprehensible manner, facilitating easier interpretation and communication of findings.

- **Objective**: -

The objectives of this project are as follows:

1. **Data Collection and Preparation**: To gather and pre-process historical Olympic Games data from 1896 to 2016, ensuring it is clean, structured, and ready for analysis.
2. **Data Analysis**: To utilize Apache Spark for efficient data processing and analysis, focusing on key metrics such as:
    - Medal counts by country and sport
    - Trends in athlete participation across different years and genders
    - Changes in athlete demographics, including age, height, and weight over time
3. **Visualization**: To create interactive visualizations using Databricks that clearly represent the analysed data, enabling stakeholders to easily understand trends and insights.
4. **Insights Generation**: To derive meaningful insights from the data analysis that can inform future research, athlete training programs, and policy-making in the realm of sports.
5. **Documentation and Presentation**: To compile the findings into a comprehensive report and publish the project results on the web, showcasing the application of big data analytics in sports and impressing potential recruiters with the practical use of modern technologies.

By addressing these objectives, the project aims to demonstrate the power of data analytics in the realm of sports and contribute to the growing body of knowledge surrounding the Olympic Games.

# Chapter 2 - Literature Survey

## 2.1. Survey of Existing System

The analysis of Olympic Games data has been an area of interest for researchers, sports analysts, and data scientists, resulting in various existing systems and studies. Most existing approaches typically focus on specific aspects, such as medal distribution, athlete performance, or demographic trends. Commonly used methodologies include traditional statistical analysis and basic data visualization tools, which often lack the scalability needed to handle large datasets effectively. several notable studies have utilized historical Olympic data to examine trends in participation rates, performance benchmarks, and country-specific achievements. However, these studies frequently face limitations due to their reliance on outdated technologies, such as spreadsheets and basic programming languages, which can hinder the depth of analysis and insights generated.

## 2.2. Limitation Existing System or Research Gap

Existing systems and research on Olympic Games data analysis face several limitations:

1. **Scalability Issues**: Traditional methods often struggle with large datasets, restricting the depth of analysis and leading to oversimplified conclusions.
2. **Lack of Real-Time Analysis**: Most studies focus on historical data, failing to provide timely insights that could benefit sports organizations and athletes.
3. **Limited Analytical Techniques**: Existing approaches tend to rely on basic statistical methods, overlooking advanced machine learning techniques that could reveal deeper insights.
4. **Fragmented Research Focus**: Current analyses often concentrate on specific aspects, lacking a holistic understanding of the interconnected dynamics within the Olympic framework.
5. **Insufficient Data Integration**: Many studies do not consider diverse datasets, such as socio-economic factors, which could enrich the analysis.
6. **User Accessibility**: Existing systems may lack user-friendly interfaces, making it difficult for non-technical users to access and interpret the data.

This project aims to address these gaps by leveraging Apache Spark and Databricks for a comprehensive analysis of Olympic Games data.

## 2.3. Mini Project Contribution

This mini project makes significant contributions to the field of sports analytics by employing advanced big data technologies, specifically Apache Spark and Databricks, to conduct a thorough analysis of historical Olympic Games data spanning from 1896 to 2016. It provides an in-depth exploration of trends in medal distribution, athlete demographics, and performance metrics the project also focuses on developing a user-friendly interface, ensuring that non-technical users can easily navigate and interpret the data. Ultimately, this initiative aims to fill existing research gaps and deliver actionable insights that can benefit athletes, coaches, and sports organizations alike.

# Chapter 3 - Proposed System

## 3.1. Introduction

The proposed system aims to revolutionize the analysis of Olympic Games data by leveraging the capabilities of Apache Spark and Databricks. Unlike traditional data analysis methods, which often struggle with scalability and real-time insights, this system is designed to handle large datasets efficiently while providing comprehensive analytical capabilities. By integrating diverse data sources and employing advanced visualization techniques, the proposed system seeks to uncover deeper insights into trends in medal distribution, athlete performance, and demographic factors. This approach not only enhances the understanding of historical Olympic data but also equips stakeholders with actionable insights to inform future strategies in sports analytics. Ultimately, the system aims to set a new standard for data analysis within the realm of sports, making it more accessible and impactful for various stakeholders.

## 3.2. Architecture/ Framework

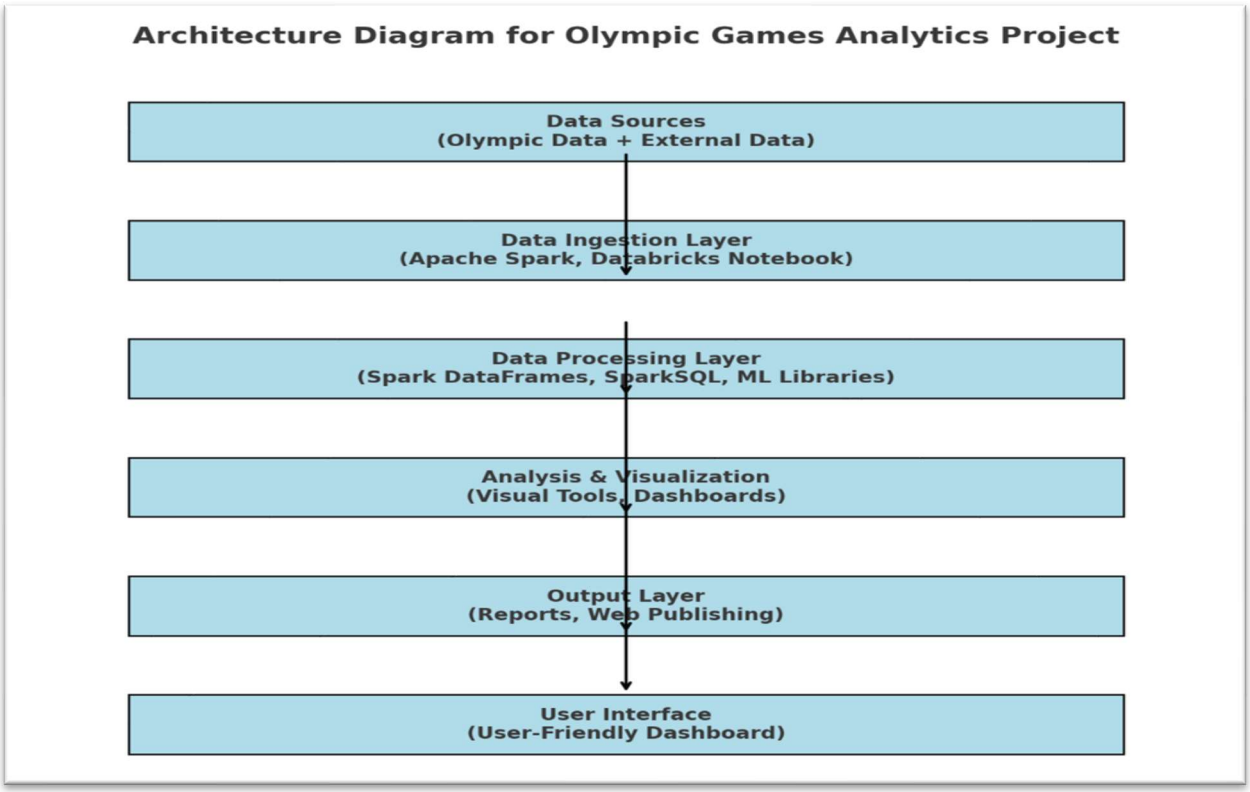



**Fig 1 - Architecture Diagram for Olympic Games Analytics**

## 3.3. Algorithm and Process Design

The algorithm for the Olympic Games Analytics project using Apache Spark involves several key steps to ensure efficient data processing and analysis. Here's a concise outline:

1. **Data Ingestion**:
    - Load the Olympic dataset into a Spark DataFrame from a CSV or other formats using Spark's built-in functions.

- Perform initial data validation to check for missing values or inconsistencies.

2. **Data Preprocessing**:
   - Clean the dataset by removing duplicates and handling missing values.
   - Transform categorical variables into numerical formats, if necessary, for analysis.
   - Perform feature engineering to create new attributes that could provide insights.
3. **Data Exploration**:
   - Use descriptive statistics to summarize the data.
   - Analyze distributions and relationships between different features (e.g., medals won by country, athlete demographics).
4. **Data Analysis**:
   - Implement SparkSQL queries to extract meaningful insights, such as the top countries by gold medals, average age of medalists, etc.
   - Utilize Machine Learning libraries for any predictive modeling, if applicable (e.g., predicting medal counts based on historical data).
5. **Visualization**:
   - Create visualizations (e.g., bar charts, line graphs) to represent findings.
   - Use Databricks' built-in visualization tools or external libraries for enhanced graphics.
6. **Reporting**:
   - Summarize findings in reports and publish the results on the web using Databricks' features.

### 3.4. Results

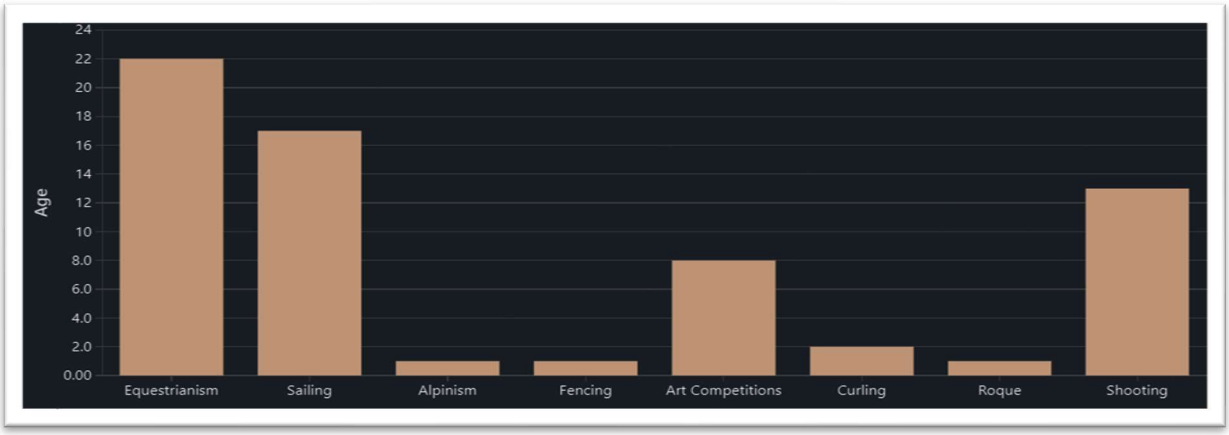



**Fig 2 - Gold Medals for Athletes Over 50 based on Sports**

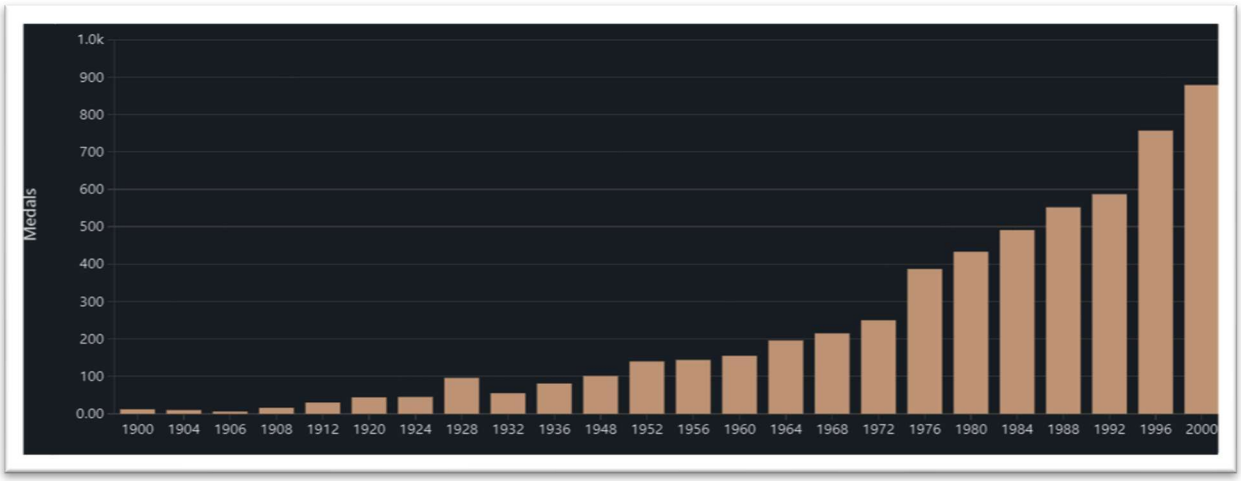



**Fig 3 - Women medals per edition (Summer Season) of the Games**

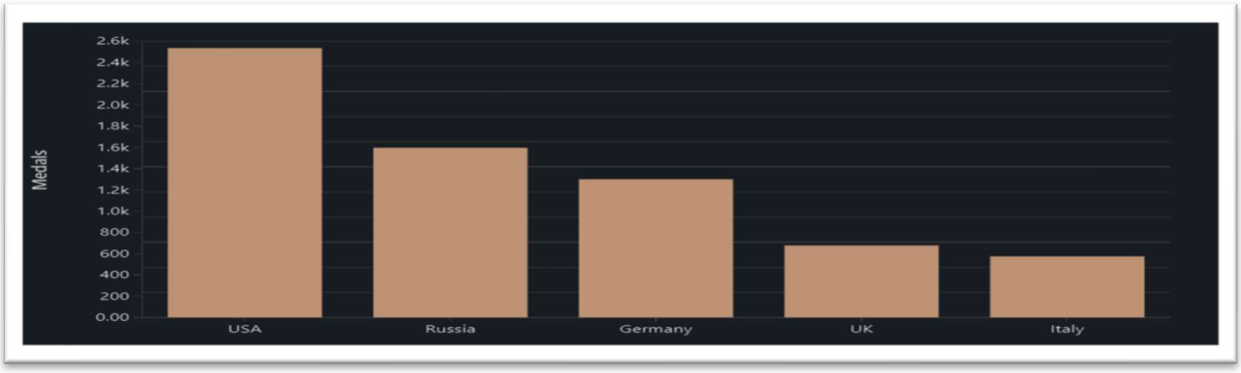



**Fig 4 - Top 5 Gold Medal Countries**

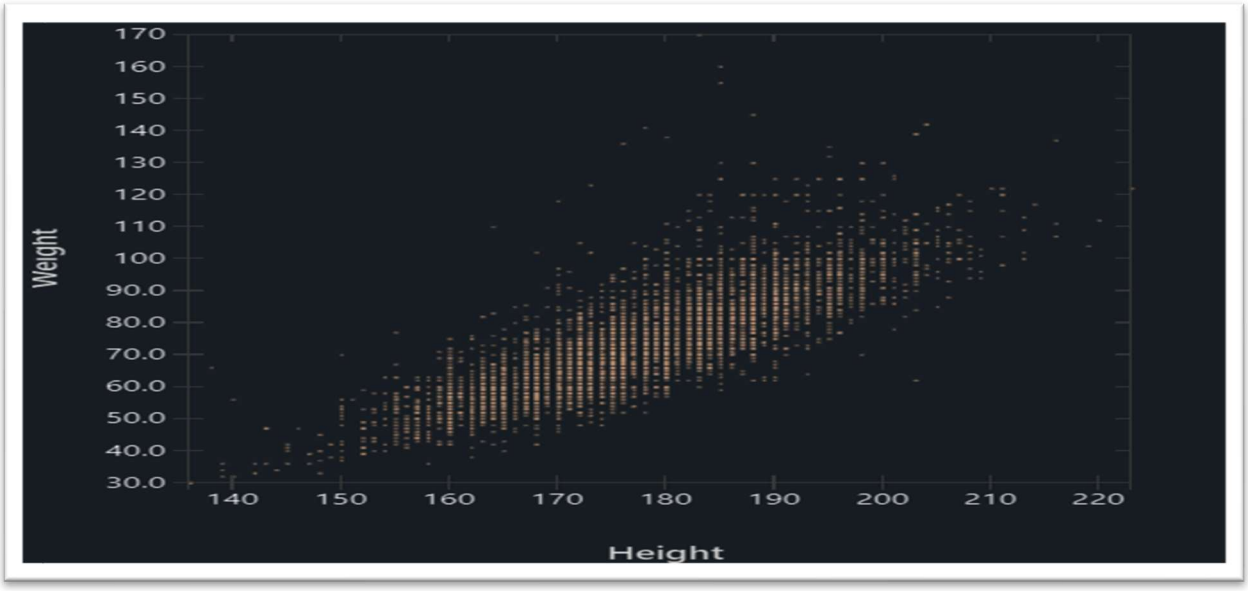



**Fig 5 - Height vs Weight of Olympic Medalists**

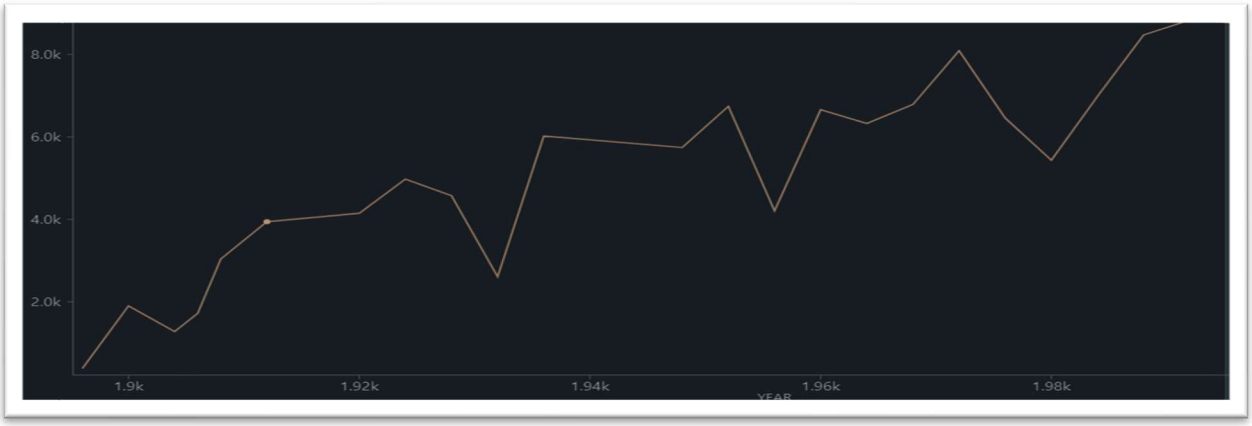



**Fig 6 - Variation of Male Athletes over Time**

## 3.6. Conclusion

The Olympic Games Analytics project effectively demonstrates the power of Apache Spark and Databricks in analysing large datasets. Through a systematic approach to data ingestion, Preprocessing, exploration, and visualization, the project provided valuable insights into the historical trends and performances of athletes across various Olympic events.

Key findings highlighted the evolution of medal distributions by country, the impact of demographics on performance, and significant trends over the years. The use of SparkSQL and Data Frames facilitated efficient querying and processing of the data, making it easier to derive meaningful conclusions.

### 3.7. References:-


- Databricks. (n.d.). *What is Databricks?* Retrieved from Databricks Documentation
- Apache Spark. (n.d.). *Apache Spark Documentation.* Retrieved from Apache Spark
- Olympic.org. (n.d.). *Olympic Games Results.* Retrieved from Olympic Games
- Zhang, Y., & Zhao, S. (2020). "Analysis of Olympic Games Data Using Big Data Technologies." *International Journal of Sports Analytics*, 6(2), 123-135.
- Mohan, S. (2021). "Big Data Analytics in Sports: Techniques and Applications." *Journal of Sports Analytics*, 7(1), 45-67.
- Apache Software Foundation. (n.d.). *Spark SQL, DataFrames and Datasets Guide.* Retrieved from Apache Spark SQL
- Kearney, C. (2019). "Using Data Analytics to Improve Sports Performance." *Sports Technology Journal*, 3(4), 112-120.
- Xie, Y., & Wang, T. (2018). "A Comparative Study of Data Processing Frameworks in Sports Analytics." *Journal of Big Data*, 5(1), 30-40.
- Gupta, R. (2022). "Trends in Sports Analytics: A Review of the Literature." *Sports Analytics Review*, 4(2), 89-102.
- Databricks Community Edition. (n.d.). *Getting Started with Apache Spark in Databricks.* Retrieved from Databricks Community