JAWAHAR EDUCATION SOCIETY'S

A. C. PATIL COLLEGE OF ENGINEERING

PLOT NO-17, SEC-4, KHARGHAR, NAVI MUMBAI - 410210
Approved by AICTE, New Delhi, Affiliated to University of Mumbai

Olympic Games Analytics Using Apache Spark

Introduction

Analyzing Olympic Games Data with Apache Spark and Databricks

01 **Project Significance**

Understanding the importance of the Olympic dataset in sports analytics.

02 **Project Focus**

Analyzing Olympic data from 1896 to 2016 using Apache Spark.

03 **Modern Tools Used**

Highlighting the role of Spark and Databricks in real-world data analysis.

04 **Motivation**

Exploring trends in sports, athletes, and country performance.

05 **Project Objectives**

Identifying trends and insights in Olympic data.

06 **Architecture Overview**

How data is processed and visualized using Apache Spark.

07 **Technologies Used**

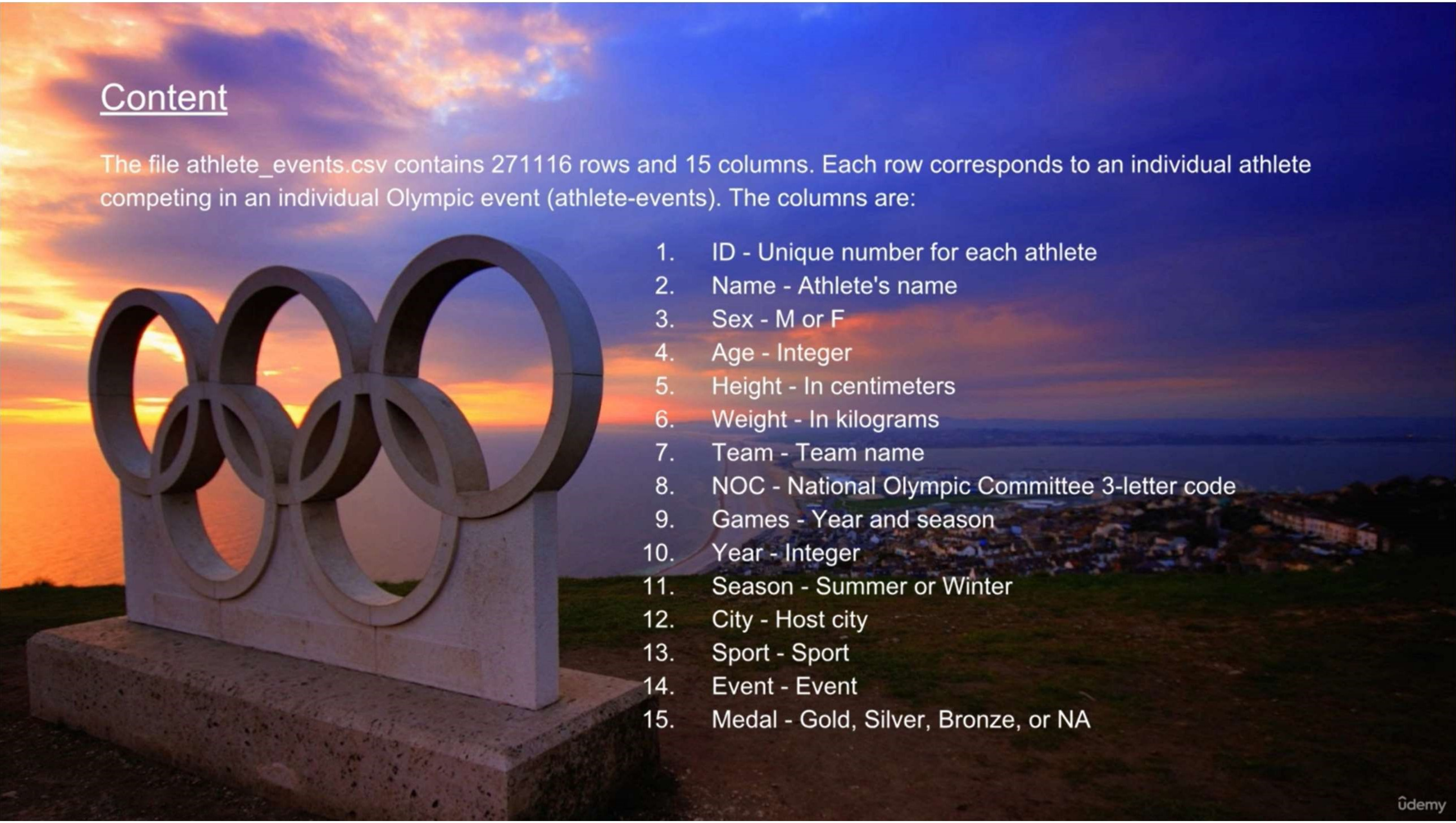
Overview of Apache Spark, Databricks, and visualization tools.

08 **Results Achieved**

Insights on medal distribution and athlete trends over time.

09 **Conclusion**

Demonstrating Apache Spark's power for large-scale analysis.



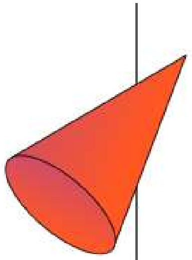
Content

The file athlete_events.csv contains 271116 rows and 15 columns. Each row corresponds to an individual athlete competing in an individual Olympic event (athlete-events). The columns are:

1. ID - Unique number for each athlete
2. Name - Athlete's name
3. Sex - M or F
4. Age - Integer
5. Height - In centimeters
6. Weight - In kilograms
7. Team - Team name
8. NOC - National Olympic Committee 3-letter code
9. Games - Year and season
10. Year - Integer
11. Season - Summer or Winter
12. City - Host city
13. Sport - Sport
14. Event - Event
15. Medal - Gold, Silver, Bronze, or NA

Motivation

Leveraging Big Data for Sports Insights



01

Rich Dataset Availability

The Olympic dataset spans from 1896 to 2016, offering extensive data for analysis.

02

Trend Analysis in Sports

Analyzing trends in sports and athlete performance can reveal critical insights.

03

Enhancing Athlete Performance

Data-driven insights can significantly improve future athlete training and performance.

04

Predicting Medal Outcomes

Predictive analytics can help forecast medal outcomes based on historical data.

05

Importance of Big Data Skills

Learning big data handling with Apache Spark is essential for aspiring data scientists.

06

Utilizing Modern Tools

Apache Spark and Databricks facilitate real-time data analysis and visualization.

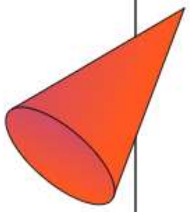
07

Scalable Data Analysis

The project leverages distributed processing power for scalable big data analytics.

Project Objectives

Analyzing Olympic Games Data with Apache Spark and Databricks



Data Exploration

Utilize Apache Spark to explore Olympic Games data efficiently and effectively.

Trend Identification

Identify trends in top-performing countries and gender participation rates over time.

Visualization Techniques

Employ Databricks for visual representation of findings, enhancing understanding of data.

Publication of Findings

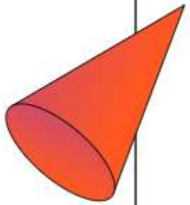
Publish insights using Databricks notebooks, showcasing results to potential recruiters.

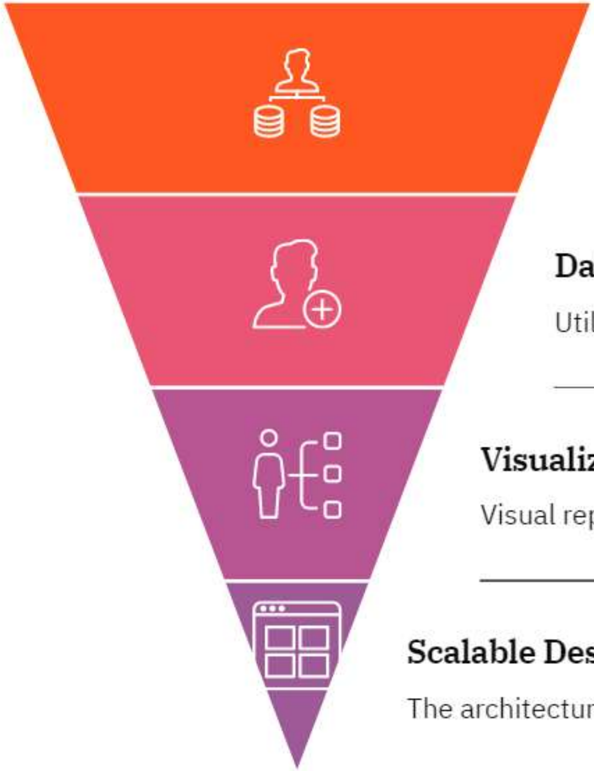
Scalable Architecture

Leverage the distributed processing power of Spark for scalable big data analysis.

Architecture Overview

Leveraging Apache Spark for Olympic Data Analytics





Data Loading

Data from the Olympic dataset is ingested into Apache Spark for processing.

Data Transformation

Utilizing SparkSQL, the dataset undergoes transformation to extract meaningful insights.

Visualization

Visual representations are created using Databricks tools and Python libraries for analysis.

Scalable Design

The architecture supports scalability, leveraging Spark's distributed processing capabilities.

Technologies Utilized

Tools for Olympic Games Data Analysis

Apache Spark

A robust framework for big data processing, facilitating large-scale analytics.

Databricks

A collaborative platform offering Spark as a service for real-time analysis.

SparkSQL

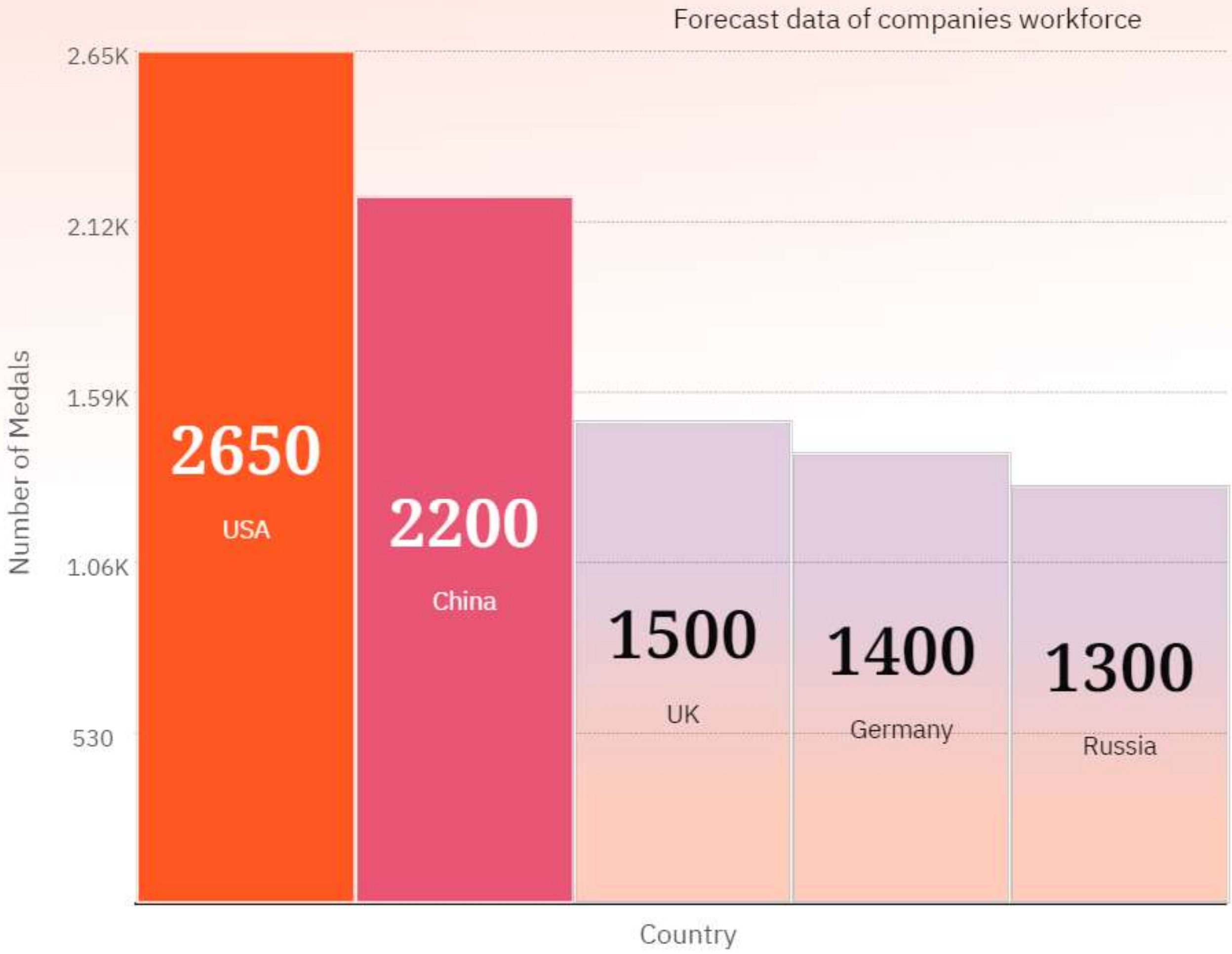
Enables structured data querying and transformation for insightful analysis.

DataFrames

Supports structured data manipulation, streamlining data handling processes.

Visualization Tools

Utilizes Databricks' in-built charts and graphs for effective data representation.



Results: Medal Distribution Insights

Analyzed medal distribution by country and gender with athlete trends.

Conclusion

Insights from Olympic Games Analytics Using Apache Spark

01

Large-Scale Data Analysis

Apache Spark effectively handles extensive Olympic Games datasets, showcasing its analytical capabilities.

02

Simplified Workflow

Databricks streamlines processes for data scientists, enhancing productivity and collaboration.

03

Valuable Insights

The analysis provides essential insights for sports analysts and recruiters regarding athlete performance.

04

Portfolio Development

The project serves as a valuable addition to data science portfolios, showcasing practical skills.