

PYSPARK

MULTI CLOUD

Project

Football Data Analysis

Batch Name:

BI-V14(AWS, Python, PySpark)

Team Members

* Romit Gupta (30014919)
* Avishek Saha (30014806)
* Manoj Choudhary (30014917)
* Saurav Rajput (30013998)
* A. Naveen Kumar Reddy (30015284)
* Anem Sai Krishna (30014916)

**Project Guide Name**

* Saransh Jain

INDEX

Topic Page no.

1. Introduction --------------------------------------------------------------- 4-7
2. Tools and Technologies -------------------------------------------------- 8
3. Problem Statement ------------------------------------------------------- 9
4. Implementation ----------------------------------------------------------- 10-15
5. Project Report Based on Business Values --------------------------- 16-19
6. Advantages of football data analytics --------------------------------- 20
7. Conclusion ----------------------------------------------------------------- 21
8. Data Source Files ---------------------------------------------------------- 21

**1. INTRODUCTION**

1.1 PROJECT OVERVIEW:

The goal of the PySpark Football Data Analysis project is to conduct a thorough analysis of football player and game data by utilizing PySpark and Apache Spark's dynamic distributed computing capabilities. With an emphasis on player performance and game statistics, this mini project investigates how big data methods might be used to extract insights from enormous datasets in the sports sector.

1.2 PROJECT OBJECTIVE:

The primary goal of this project is to analyze and derive meaningful insights from two distinct datasets: one containing detailed player information and the other tracking game-level data. By performing exploratory data analysis (EDA), aggregation, and various transformations, we seek to uncover patterns and trends that can help us understand player performance metrics, game outcomes, and other key factors influencing football matches.

1.3 DATASETS :

Table – 1:

|  |  |
| --- | --- |
| **Players\_Src1** | |
| Year | Integer |
| Month | String |
| Match | String |
| Players Name | String |
| Club Name | String |
| Country Name | String |
| Position | String |
| Jersey Number | Integer |
| D. O. B. | String |
| Nationality | String |
| Appearances | Integer |
| Goals Scored | Integer |
| Goals Assist | Integer |
| Total Shots | Integer |
| Shots on target | Integer |
| Fouls made | Integer |
| Fouls suffered | Integer |
| Yellow Card | Integer |
| Red Card | Integer |
| Goals saved | Integer |
| Goal Conceded (Stopped) | Integer |
| Total Penalty | Integer |
| Successful Penalty | Integer |
| Salary | String |

Table – 2:

|  |  |
| --- | --- |
| **Matches\_Src2** | |
| Year | Integer |
| Match name | String |
| League name | String |
| Club Name | String |
| Coach | String |
| Country name | String |
| Manager | Integer |
| Owner | String |
| Appearances | Integer |
| Wins | Integer |
| Losts | Integer |
| Drawn | Integer |
| Clean Sheets | Integer |
| Net worth | Integer |

Table – 3:

|  |  |
| --- | --- |
| **DIM.Date\_Src3** | |
| Date Key | Integer |
| Full date | Timestamp |
| Day of week | Integer |
| Day num in month | Integer |
| Day num overall | Integer |
| Day name | String |
| Day abbrev | String |
| Weekday flag | String |
| Week num in year | Integer |
| Week num overall | Integer |
| Week begin date | Timestamp |
| Week begin date key | Integer |
| Fb month | Integer |
| Month num overall | Integer |
| Month name | String |
| Month abbrev | String |
| Quarter | Integer |
| Fb year | Integer |
| Yearmo | Integer |
| Fiscal month | Integer |
| Fiscal quarter | Integer |
| Fiscal year | Integer |
| Last day in month flag | String |
| Same day year ago date | Timestamp |

**2. TOOLS AND TECHNOLOGIES**

2.1 ENVIRONMENT SETUP FOR MINIPROJECT

* Software:
  + AWS account with S3 access
  + Microsoft Edge / Google Chrome
  + Data Bricks Account with required privileges

3. PROBLEM STATEMENT

**3.1 Objective:**

Development of a Analysis & Reporting System of Football Game

**3.2 Abstract of the project:**

This project shows how to analyze massive amounts of football data using Apache Spark and its Python API, PySpark. Analyzing two datasets—one that tracks individual match performance statistics and the other that has comprehensive player information—is the main goal. Traditional methods frequently find it difficult to efficiently process and derive valuable insights from the increasing number of football data. Thus, this project uses Python 3.9.11's user-friendliness and Spark 3.5.3's distributed computing capability to process, transform, and analyse enormous amounts of data.

The project starts by loading raw data from CSV files that include match information and player statistics. Data preprocessing operations including cleaning, addressing missing values, and combining the datasets are completed effectively with PySpark's robust DataFrame API. To find key performance indicators (KPIs) and relationships between player statistics and game results, such as goals scored, assists, and team performance, exploratory data analysis, or EDA, is used.

The project also investigates more complex analysis, like combining player and team-level data, comparing performances, and utilizing Matplotlib and Seaborn to visualize trends. Spark's scalability enables the analysis to manage bigger datasets, guaranteeing that the method can be expanded to more intricate and substantial data in the future. In addition to showcasing PySpark's capacity to grow and handle enormous datasets, this application shows how contemporary big data tools may improve sports analytics by offering insightful information about player performance, team tactics, and game results.

## **3.3 Technology use:**

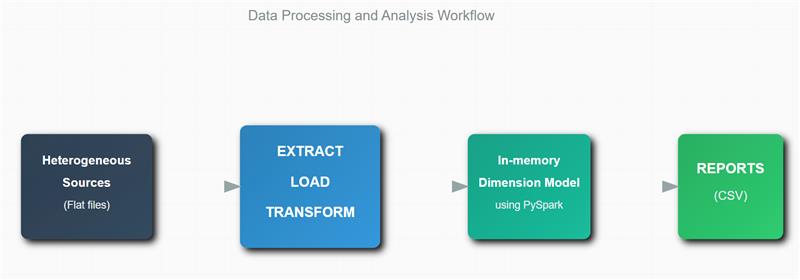
* Python
* Jupyter Notebook
* Pyspark

4. IMPLEMENTATION

4.1 Summary of the functionality to be built:

The functionality to be built in this project involves creating an end-to-end data pipeline that processes, cleans, analyzes, and visualizes large football datasets. By leveraging PySpark's distributed computing capabilities, this project aims to provide actionable insights into player performance, team strategies, and match outcomes while ensuring that the solution is scalable and can handle large datasets efficiently. The visualizations and reports generated will offer valuable decision-support tools for football managers, analysts, and enthusiasts.

## 4.2Guidelines on the functionality to be built:

**Project flow :**

**4.2.1. Heterogeneous Sources:**

Data Sources: Data comes from  CSV files(match.csv  , player.csv). These sources may have different formats, structures, and technologies which are uploaded in AWS S3 bucket.

**4.2.2. ELT (Extract, Load, Transform):**

* **Extract:** Data is pulled from these heterogeneous sources.
* **Load:** The raw data is then loaded into a various dataframes. This dataframes enables efficient processing and analysis. At this stage, no transformation occurs yet, as the focus is on gathering all necessary data.
* **Transform:** Data is then transformed using PySpark for cleaning and structuring.   
  Example - Removing duplicates or correcting missing values.

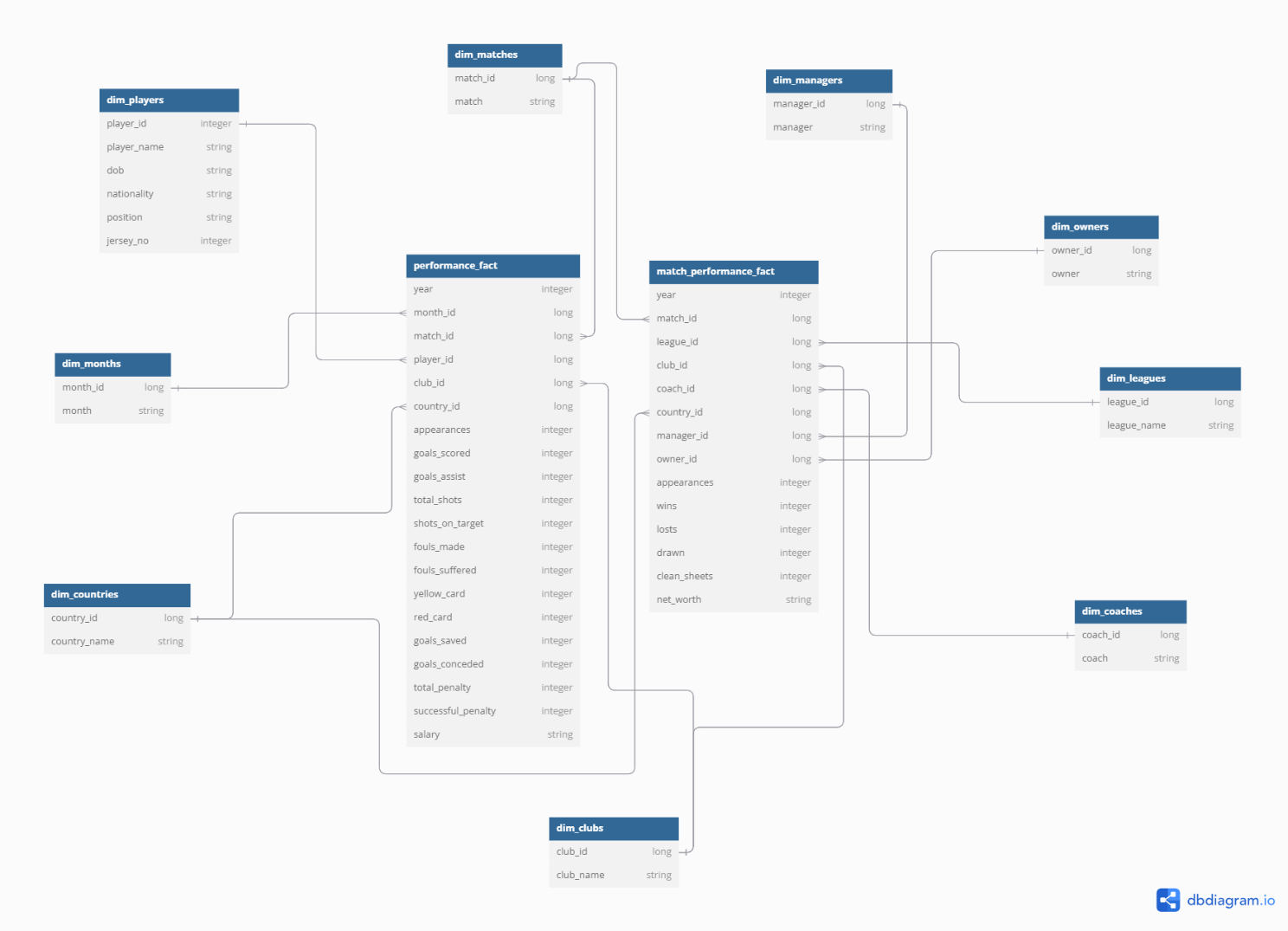
**4.2.3. In-memory Dimension Model using PySpark:**

* After the raw data is loaded and transformed, it is stored in an in-memory dimension model. This model is typically structured in a galaxy schema , where:
  + **Facts:** Quantitative data (e.g., players\_performance\_fact , match\_performance\_fact).
  + **Dimensions:** Descriptive attributes (e.g., players, months, countries, clubs etc.).
* By using PySpark, this data model is built in memory, which allows for fast access and analysis. Spark’s distributed computing capabilities help in processing and managing large datasets efficiently for analytics.

**4.2.4. Reports (CSV)**:

* The final step is generating **reports** in the form of **CSV files**. These reports summarize key insights from the football data analysis. For example:
  + Player performance analysis (goals, assists, average match ratings).
  + Team rankings based on performance metrics.
* The transformed data is written to CSV files, which can then be used for further analysis, shared with stakeholders, or visualized and also can be saved in AWS S3.

In-Memory dimensional model relationship for reference purpose.

****

**Dimension Tables:-**

In the context of PySpark and database systems in general, a dimension table is a table used in data warehousing and analytics to store descriptive or categorical information that is often used for filtering, grouping, or analyzing fact data (usually stored in fact tables).

Dimension tables in this project:-

* dim\_players-

1. player\_id (int)
2. player\_name (string)
3. dob (string)
4. nationality (string)
5. position (string)
6. jersey\_no (int)

* dim\_months-

1. month\_id (int)
2. month (string)

* dim\_countries-

1. country\_id (int)
2. country\_name (string)

* dim\_clubs-

1. club\_id (int)
2. club\_name (string)

* dim\_coaches-

1. coach\_id (int)
2. coach (string)

* dim\_leagues-

1. league\_id (int)
2. league\_name (string)

* dim\_owners-

1. owner\_id (int)
2. owner (string)

* dim\_managers-

1. manager\_id (int)
2. manager (string)

* dim\_matches-

1. match\_id (int)
2. match (string)

**Fact Tables:-**

In the context of PySpark and data warehousing, a **fact table** refers to a central table that stores quantitative data for analysis. Fact tables typically contain facts (measurable, numerical data) and foreign keys that reference dimension tables, which contain descriptive attributes.

Fact tables in this project:-

* players\_performance\_fact-
  1. year (int)
  2. month\_id (long)
  3. match\_id (long)
  4. player\_id (long)
  5. club\_id (long)
  6. country\_id (long)
  7. appearances (int)
  8. goal\_scored (int)
  9. goals\_assists (int)
  10. total\_shots (int)
  11. shots\_on\_target (int)
  12. fouls\_made(int)
  13. fouls\_suffered (int)
  14. yellow\_card (int)
  15. red\_card (int)
  16. goals\_saved(int)
  17. goals\_ conceded (int)
  18. total\_penalty (int)
  19. successful\_penalty (int)
  20. salary (int)
* match\_performance\_fact-

1. year (int)
2. match\_id (long)
3. league\_id (long)
4. club\_id (long)
5. coach\_id (long)
6. country\_id (long)
7. manager\_id (long)
8. owner\_id (long)
9. appearances (int)
10. wins (int)
11. losts (int)
12. drawn (int)
13. clean\_sheets (int)
14. net\_worth (string)

**Data Transformation for data warehouse:**

1. Load the Data into Dimension tables using the CSV files provided.
2. Load the Data into Fact tables using the CSV files provided.
3. Populate Dimension tables before fact table
4. Dimension table should have unique values with unique system generated IDs. Rank() function can be used to generate the unique ID (sequence). Same for Fact ID
5. Other ID columns in Fact table example in Players Fact table – Match\_type\_id, player\_id, club\_id etc. column should be populated with id columns values from corresponding dimension table

**5. PROJECT REPORT BASED ON BUSINESS VALUES**

Business Values based on the dimension and fact tables:-

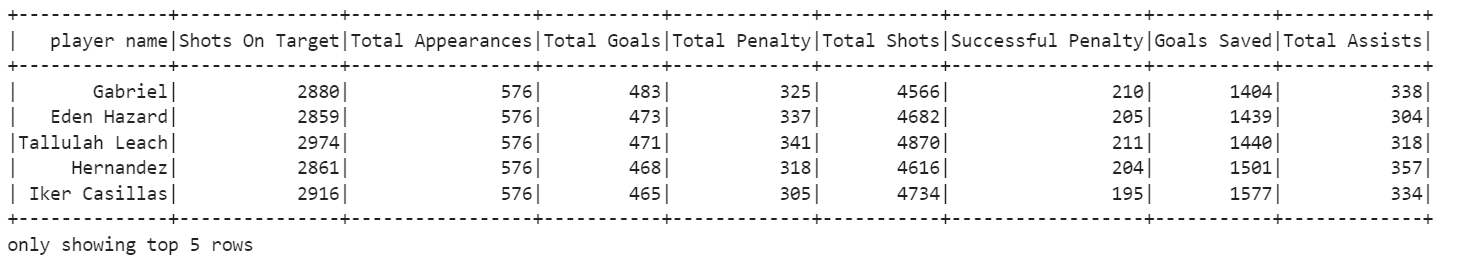
**Player performance analysis**

**Identify Top Players by Goals and Assists(Helpful for football clubs)**

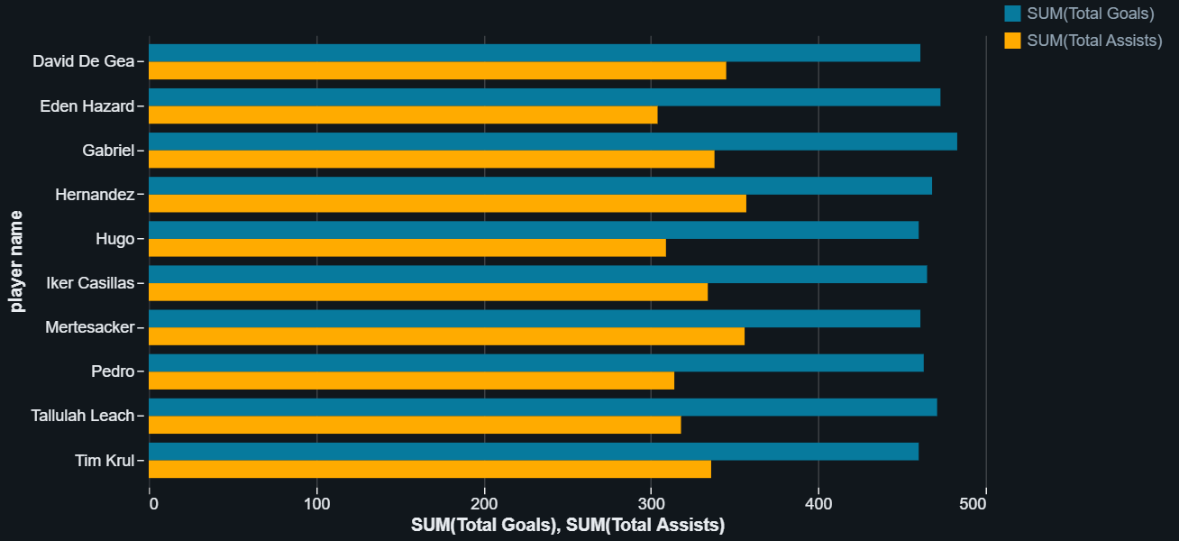
By examining player statistics (goals scored, assists, fouls), clubs can identify top-performing players and make informed decisions regarding contracts, buying or selling, player development.

**Populating Sports websites**

Sports websites can populate these datas in their platform to enhance the quality of football datas to the end users.



Player Report Visualization :-



**Find who has maximum appearances in International matches**

**Brand Ambassadorship & Endorsements**

**Personal Brand:**

An experienced football player often has a strong personal brand, built over years of playing at a high level. This brand can be leveraged for endorsements, sponsorships, and partnerships with companies, making them valuable assets for brands looking to reach specific consumer segments.

**Influence and Reach:**

Players with a large following on social media platforms can be influencers for brands, helping them engage with fans and customers on a personal level.

**Product Partnerships:**

Football players often collaborate with companies for product endorsements in sectors like apparel, footwear, health supplements, or even technology.

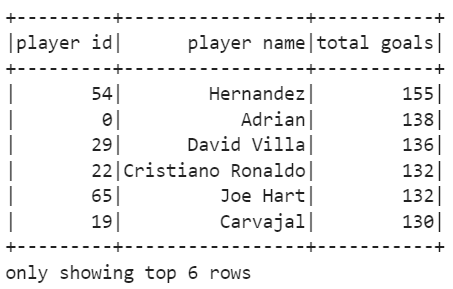
A black and white text

Description automatically generated with medium confidence

**find top goal scorers of a specific year**

**Sports website data**

Sports websites can use the data to populate their information as much as possible as the old football data are not that much available in the market.



Visualization of top goal scorer of 2010:-

A graph with red line

Description automatically generated

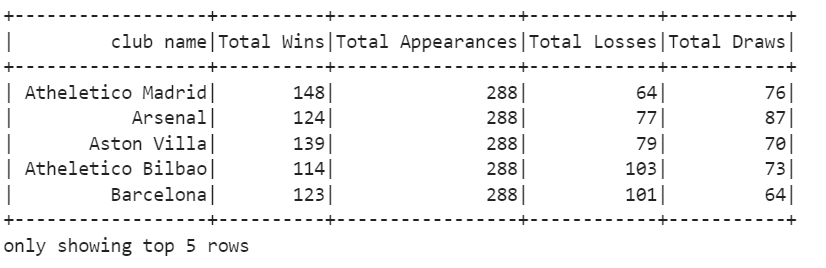
**Club report of Total Wins , Total Appearances ,Total Losses ,Total Draws**

**League and Competition Insights**

Clubs can compare their performance against league averages or top performers, providing insights into areas for improvement. Trend Analysis: Longitudinal analysis of match results and player performances over time can reveal trends that influence future strategies.

**Ownership and Sponsorship**

The top companies can fund or sponsor on the clubs based on their performance



Visualization of Clubs’ Report :-

A graph with multiple colored bars

Description automatically generated with medium confidence

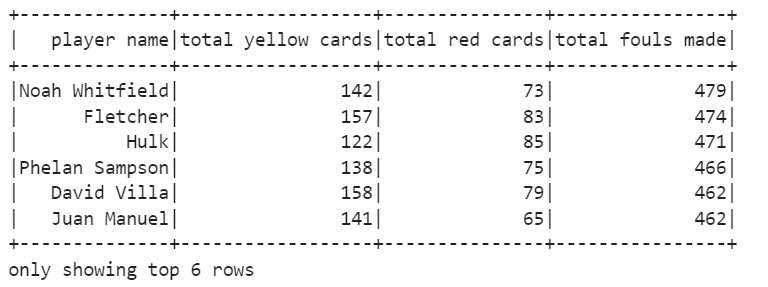
**Player Injury & Suspension Analysis**

**Helpful for club**

By tracking fouls committed, yellow/red cards (for goalkeepers), clubs can better understand a player's tendency for suspensions. This information is vital for managing player health, avoiding suspensions, and understanding risks. The clubs can also avoid to extend contract with those player who have highest tendency to get a card.

**Helpful for sports website**

They can populate these datas also in their website

****

**Visualization of red card , yellow card and total foul made for the players :-**

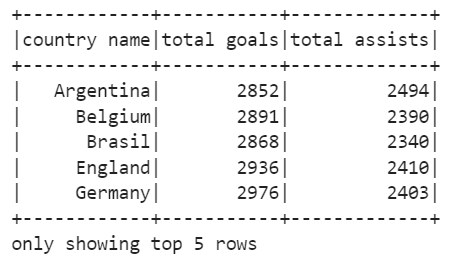
**A graph with numbers and text

Description automatically generated with medium confidence**

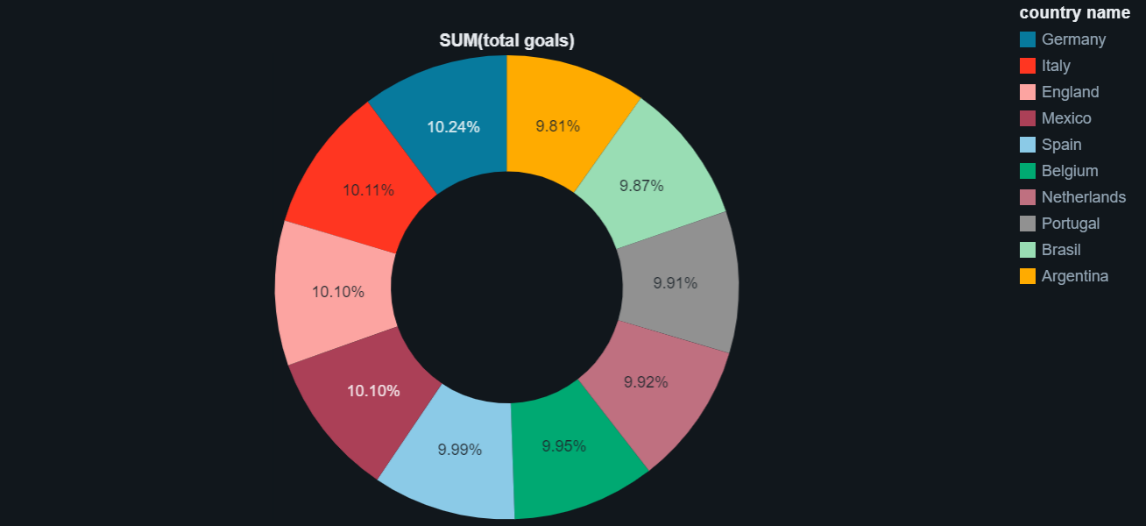
**Country-Specific Performance Insights**

**Scouting project insights**

By getting the country specific statistics the club can build a scouting project on those countries which are performing in the club levels as well as the international level.



Visualization of country based performance report :-



6. **Advantages of Football Data Analysis**

6.1. **Improved Team Performance**  
**Tactical Insights:** A detailed analysis of a team’s performance can uncover weaknesses in tactics, positioning, and decision-making. Identifying these areas can help coaches fine-tune strategies and improve gameplay.  
**Match Preparation:** By analysing past games and opponents, teams can develop more effective game plans tailored to their next opponents' strengths and weaknesses.  
**Player Development:** Performance analysis tools can track player movement, pass accuracy, decision-making, and more. This allows for personalized feedback and targeted skill development.

6.2. **Injury Prevention & Player Welfare**

**Monitoring Player Injuries:** Through analysis of player movements and physical stress (such as foul committed , file conceded ), coaches can monitor to prevent injuries.  
**Recovery Strategies:** Analysis can help track recovery times and injury rates, enabling more effective management of rest periods and recovery techniques.

6.3. **Fan Engagement & Media Content**

**Interactive Visualizations:** Analysis can produce visual content such as heat maps, player tracking, and tactical breakdowns, which can be used to engage fans and audiences. This provides an added layer of entertainment and insight, increasing fan investment in the game.  
**Media Coverage:** Analysts and commentators use statistical breakdowns and tactical analysis to provide more insightful coverage during live broadcasts or in post-match discussions.

6.4. **Strategic Planning and Long-Term Vision**

**Trend Identification:** Football analysis helps teams understand long-term trends (such as changes in playing styles, tactics, or formations) and how the game is evolving. This allows organizations to plan for future seasons and adapt to changes in the sport.  
**Club Development:** By analysing various aspects of club performance, from youth systems to senior teams, clubs can identify areas for growth, whether it's improving youth development programs, strengthening recruitment strategies, or boosting financial success through better performances.

7. **Conclusion**

Football analysis is not just for professional clubs or analysts; it's a powerful tool for enhancing all aspects of the game, from player development and tactical improvements to fan engagement and financial success. It allows for a more detailed, objective, and strategic approach to understanding the game, leading to better decisions and enhanced performances on and off the field.

8. **Data Source Files**

