Soccer Analysis: Leveraging Data for Comprehensive Innovation

DATA 225 : Database Systems for Analytics

Submitted by (Group-4):
Anitha Balachandran
Aradhya Alva Rathnakar
Bhavan Kumar Basavaraju
Mahamaya Panda
Shashi Kumar Kadari Mallikarjuna

***Abstract*—If you were to inquire with any athlete or sports team globally about what motivates them, the majority would likely respond with 'Winning'. When asked what factors would make the team win, each person had a different perspective. When we narrow down the scope of the audience to soccer, diverse opinions on the capabilities and chances of winning are being procured. Due to the large set of influencing factors, prediction is the hardest part. For many years, people believed in predictions by pundits or even an octopus in the 2010 world cup, which is unrealistic and unreliable. One needs analysis to make a proper and even adequate prediction. This project aims to expand the range and influence of analysis by providing accessible and meaningful information to soccer fanatics to make predictions and help team management examine their strengths and shortcomings. It briefly describes our project and uses the concepts of databases, ETL, and data warehousing to process the data for analysis. We analyze all this meaningful information obtained from the database, depicting how a team approaches different matches, player impacts, etc. We aim to find key factors driving innovation in soccer regarding data analysis, ultimately enabling high-quality tactical strategies and predictions before deciding by a manager, managing company, team players, wager candidates, or even fanatics.**

***Index Terms*—SoccerAnalysis, Data modeling, Data cleansing, OLTP, Dimensional Modeling, OLAP, ETL, Data warehouse, Data Visualization**

I. Introduction

SOCCER is a passion shared by millions of fans, players, and managers worldwide, with winning being the ultimate goal. However, predicting match outcomes is difficult due to the many factors involved. Traditional methods, such as relying on pundits or psychic predictions, are unreliable and fail to account for the game's complexity. Our team aims to use a data-driven approach to analyze soccer data and evaluate team performance.

II. Project Goals

1. To provide accessible and meaningful information to soccer enthusiasts to help team management examine their strengths and shortcomings.

2. To design a robust data model that effectively represents soccer data entities, attributes, and relationships to enable efficient querying and analysis.

3. To clean and transform soccer data using ETL tools like Azure Data Factory and load it into a data warehouse like Azure SQL pool on the cloud for historical analysis and reporting.

4. To use reporting tools like Power BI and SSAS to create visualizations that enhance the user experience in analyzing the information.

III. Motivation

Data collection is a tool that has enhanced the technological world and opened the scope for a wide range of analytics implementation. After the concept of modern soccer came into existence in the late 18th century, it dominated the sports world. The interest in soccer is increasing for several reasons, including placing wagers and pride in winning. These reasons have become a crucial part of soccer, and people are trying to find ways to succeed. This is where analysis and prediction play a pivotal role. The analysis gives scope to both team and manager to assess the performance, which further helps predict a match result which takes us back to supporting the first reason. Since analysis plays a massive role in soccer, we devised an idea to enhance the existing systems by building processes to collect data from various sources and provide resultant functional information using reports. All these analyses and predictions will improve and give a different perspective on how audiences and managers view the game.

IV. Literature Survey

Analytics is being developed and revolutionized continuously per changing needs and requirements. Therefore, we built our project idea on a few concepts from fast-paced data-driven environments. Mentioned below are a few concepts we referred to:

[1] Alejandro Benito Santos, Roberto Theron, Antonio Losada, Jaime S. Sampaio, and Carlos-Lago-Penas took the bar up in 2018 by contemplating dynamic variables in the pilot tool and using real-time data using live satellite feed. Player positional data and collective behavior analysis are displayed through heat maps and line graphs, making it easier to derive meaningful information.

[2] Tushar Joshi, in 2019 articulated how a data-driven approach has changed the modern game. He spotlights the importance of a player's fitness and analyzes the player's fitness level by obtaining data from wearable chips during training and matches. Such data can help calculate the type of intensity and hours needed for training, diet corrections, and proper medical conditioning.

[3] Keisuke Fujii took the analysis deeper by considering team behavior as an impacting factor. Parameters such as higher-order interactions, cognition, and body dynamics were considered and used for analysis. The survey follows two main approaches : (1) extracting easily interpretable features or rules from data and (2) generating and controlling behaviors in visually-understandable ways.

[4] Naeer Amin 2022 discussed different approaches used by other teams for optimal squad selection match analysis. Each player is evaluated through the Goal Impact metric, but the drawback was that players off-ball were not considered. Hence came real-time data feed where data is taken of player runups, the ball passes, sprint duration, and kick accuracy in calculating GIM. Big data is considered to combine past data with a live feed for better accuracy.

[5] Zachary Wu and Vince Pulido got an edge over modern analysis by determining how many opponents' difficulty impacted individual player performances. Points vs. Fixture scatterplot was created for Premier League soccer players, and 'r-squared' values indicated how much a problematic fixture moved a player's performance. The plot gave valuable insights into which player has to be targeted or avoided.

V. Project Overview and Architecture

A. *Overview*

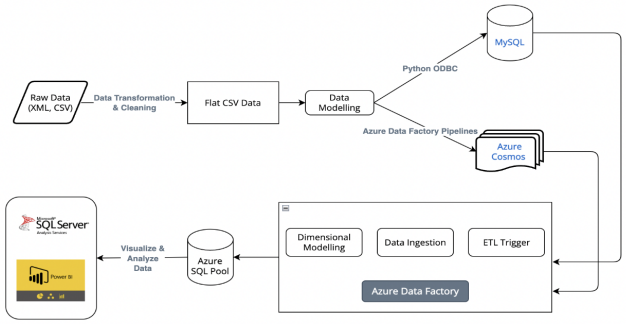


Fig. 1. Project Architecture

B. *Methodology*

*1) **Project Flow**:*

(i) For soccer data analysis, multiple datasets are collected from diverse sources, such as FIFA, Kaggle, etc. These datasets contain historical soccer data filtered primarily based on key metrics like 'Home and Away matches statistics', 'player stats,' etc. which are meaningful and helpful in performing an adequate analysis.

(ii) Based on data modeling, we cleaned the source data using Python pandas libraries and loaded the cleansed data into MySQL.

(iii) The player CSV file was dropped in the Azure storage account, from which it was loaded into CosmosDB using the Azure data factory.

(iv) We then performed dimensional modeling to decide on the granularity of the data, dimensions, and facts. Then we created ETL pipelines using Azure Data Factory to load the data into the data warehouse, i.e., Azure dedicated SQL pools on the cloud for historical analysis and reporting.

(v) This allowed for efficient data analysis using Microsoft SQL Server (SSAS) and generated valuable insights using T-SQL queries and Power BI to look at trends and patterns in data. We were able to manage and analyze large datasets for improved outcomes. Different visualizations are designed per the viewer's requirements, which in this case are soccer enthusiasts and team managers.

*2) **Database Implementation**:* All tables except the match table were in 3NF. The match table had repeating groups, which violates the 3NF rule. The repeating groups in the match table had values like shoton, shotoff, goals, penalties, etc. To bring the match table to 3NF, we moved the repeating groups' column into a new table called match_attributes using match_api_id as the foreign key column to establish a relationship between the original match table and the new table. The new table includes columns like match_api_id, shoton, shotoff, goals, subtype, etc. By splitting the match table into a new table, we eliminated the repeating groups, and all tables in the database were in 3NF ensuring data integrity. Thus, any queries that run against the database will return accurate results.

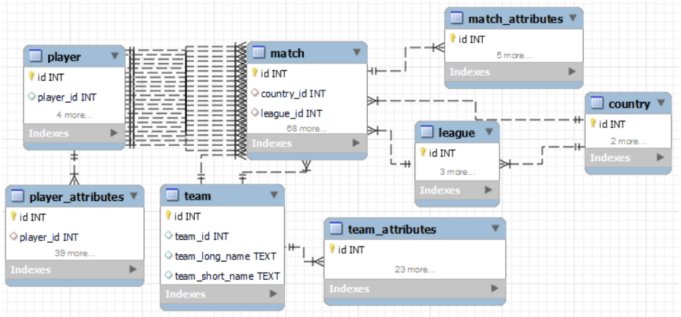


Fig. 2. Entity Relationship Diagram

*3) **About the Data**:* During the data pre-processing stage, we encountered a column in our soccer dataset containing XML data that needed to be parsed to extract relevant information for our analysis. We used pandas and XML. Tree.ElementTree libraries in Python to perform the parsing operation, extract the necessary tags and convert them into separate columns.

After parsing the data, we cleaned the null values from the dataset to ensure that the data was consistent and accurate. Finally, we saved the cleaned and transformed data into a separate CSV file named "match_attributes."

Data Cleaning:

(i) Before Data Pre-Processing:

(ii) After Data Pre-Processing:

id	match_api_id	goals	shoton	penalties	subtype
1729	489042	1	1		header
1729	489042	1	1		shot
1730	489043	1	1		shot
1731	489044	1	1		distance
1732	489045	1	1		shot
1732	489045	1	1		shot
1732	489045	1	1		volley
1733	489046	1	1		header
1733	489046			1	
1733	489046	1	1		shot
1733	489046	1	1		header
1733	489046	1	1		shot
1733	489046	1	1		shot
1734	489047	1	1		distance
1734	489047	1	1		direct_freeki
1734	489047	1	1		header
1734	489047	1	1		shot
1734	489047	1	1		shot
1735	489048	1	1		shot
1735	489048	1	1		shot
1736	489049	1	1		distance
1736	489049	1	1		header
1736	489049	1	1		header
1736	489049	1	1		header
1737	489050	1	1		header
1737	489050	1	1		distance
1737	489050	1	1		shot
1738	489051	1	1		shot
1738	489051	1	1		header
1738	489051			1	
1738	489051	1	1		distance
1739	489132	1	1		shot
1739	489132	1	1		shot
1740	489133	1	1		distance
1740	489133	1	1		header
1740	489133	1	1		header
1740	489133	1	1		shot
1740	489133	1	1		shot

*4) **Insert data into OLTP System**:*

(i) The Python script was written for MySQL to load the data from CSV files on the local system.

(ii) Azure data factory pipeline to load data from CSV file in Azure storage account to CosmosDB like below:

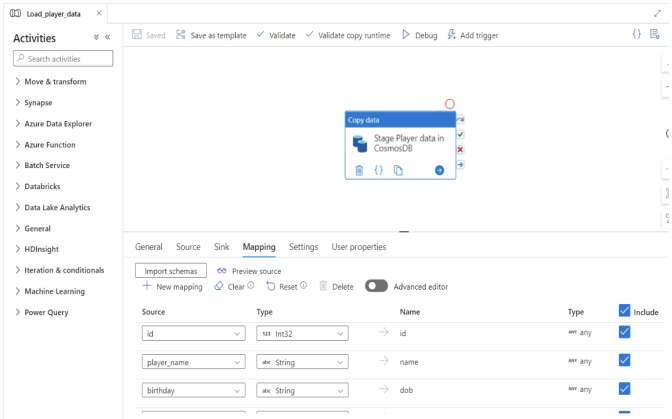


Fig. 3. Azure Data Factory

Below is a picture to show how the data looks in Cosmos DB after insertion.

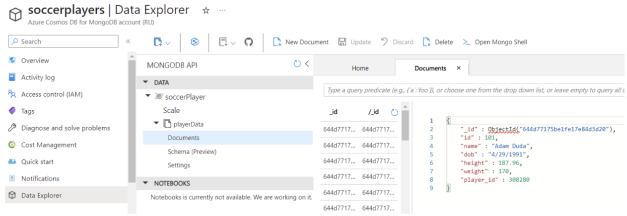
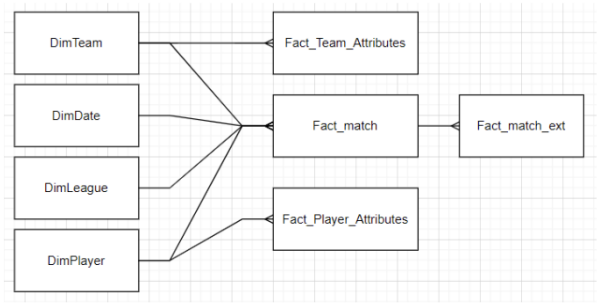


Fig. 4. Azure Cosmos DB

*5) **Dimensional Modeling**:* We decided on the dimension and fact tables based on the analysis requirements. We chose the granularity of the data to be stored as the match-level data. There were tables like Country in the OLTP database, but based on our requirement, we combined them into the DimLeague table, which helps with faster data retrieval. We also created a DimDate table for time series analysis to see historical trends and perform aggregations over time. We created a data warehouse using Constellation schema based on our analysis requirement.



Above is a dimensional modeling ER- diagram to show the relationship between different dimensions and fact tables

*6) **Data Warehouse Implementation**:* After the ETL pipelines have successfully run, the data is ingested into Azure SQL Pool which is used as the Data Warehouse for the Soccer data. Azure SQL Pool is a massively parallel processing (MPP) cloud-based data warehousing solution for large-scale data. It provides the flexibility of scale to accommodate changing workloads, and its serverless nature eliminates the need for infrastructure management. Azure SQL Pool also includes built-in machine learning features and query optimization for faster data processing. The data is ingested into Azure SQL Pool according to the developed data warehouse model, including an event table as a fact and a group table as a dimension.

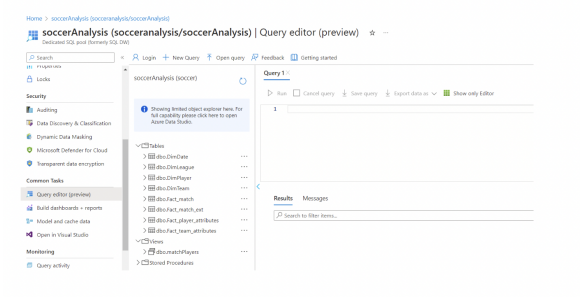


Fig. 5. Azure SQL Pool

*7) **ETL Process (Extraction, Transformation, and Loading)**:* After analyzing the soccer dataset in the OLTP and defining the necessary transformations to prepare the data for loading into the Azure SQL Pool (Data Warehouse) using Azure Data Factory, data enrichment, and data aggregation were needed. We combined the league and country OLTP table into a dimension table called DimLeague. Once the data transformation was decided, ETL was performed based on the business requirements to populate the dimensions and facts.

(i) **ETL Processes:** These are implemented with the help of Azure Data Factory. Using the native connectors of Azure Data Factory, data was extracted from various sources, including MySQL and Cosmos databases, and was transformed using built-in data flow transformations like copy data activity where we had to write queries to transform the data at the source using join statements as required. MySQL database, one of our data sources, was hosted on the local system for which a self-hosted integration runtime had to be downloaded and configured to act as a gateway to the cloud. Once the data sources and the data warehouse were configured and connected to Azure Data Factory using Linked Services, the copy data activity was used where source, destination, and mapping of columns could be done. Finally, SQL queries were used to read the transformed data from the source. The transformations included selecting only the necessary columns from the source tables, joining tables like country and league to load into a single DimLeague table, handling data type mismatches, etc. Below is a picture of a pipeline to load data into DimLeague.

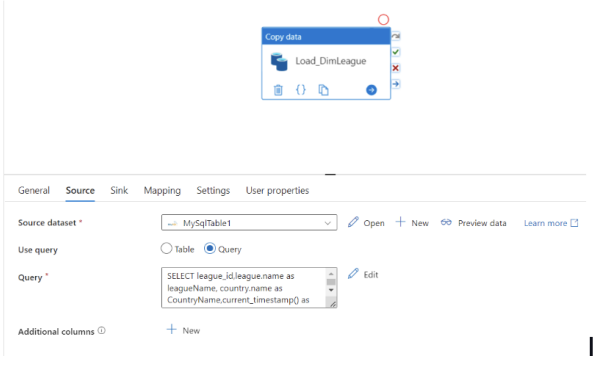


Fig. 6. Data Load to Table DimLeague

The transformed data were then loaded into destination tables, which in our case, were the dimension and fact tables in Azure dedicated SQL pools. The copy data activity let us map the source and destination table columns, which helped smooth data integration. Below is a picture of the master pipeline in Azure Data Factory that was used to load the data from OLTP to OLAP, which was run every night to ensure the data was up to date the next day.

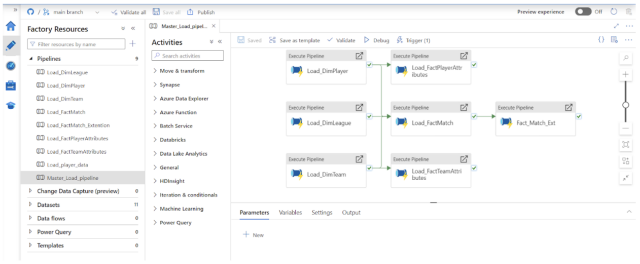


Fig. 7. Azure Data Factory cntd.

(ii) **ETL Monitoring:** The master pipeline was scheduled to run every night to ensure the up-to-date data was available for the following day to perform analysis. The pipeline runs could be monitored, and the ETL processes could be fixed based on that information.

A picture of the ETL monitoring can be found below for a better understanding.

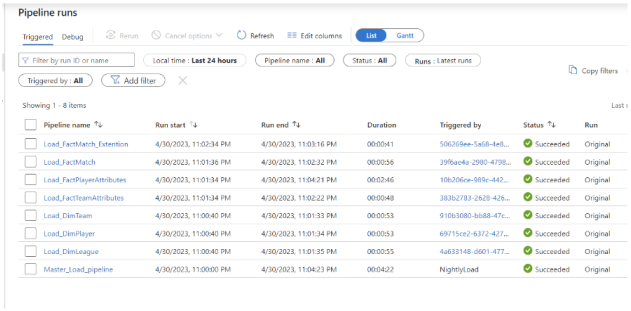


Fig. 8. ETL Pipelines running in Azure Data Factory

VI. Agile / Scrum Methodology

1. Trello

Our team decided to use Trello for our project management as it provided an easy-to-use platform that allowed us to organize and track our progress during each sprint. Throughout the project, we had a total of 8 sprints(1 Sprint = 1 week), each with specific goals and deadlines. At the beginning of each sprint, we created new sprints on the Trello board and listed the tasks that needed to be completed. We then assigned tasks to specific team members, set completion deadlines, and monitored each task's progress using the Trello board.

During each sprint, we had regular meetings to discuss the progress made on each task and to make any necessary adjustments to the project plan. These meetings were typically held via Zoom, and we also used a shared document to keep track of meeting minutes and action items. Overall,

using Trello for our project management allowed us to stay organized, track our progress, and collaborate effectively as a team. It was an important tool that helped us complete the project within the given timeline.

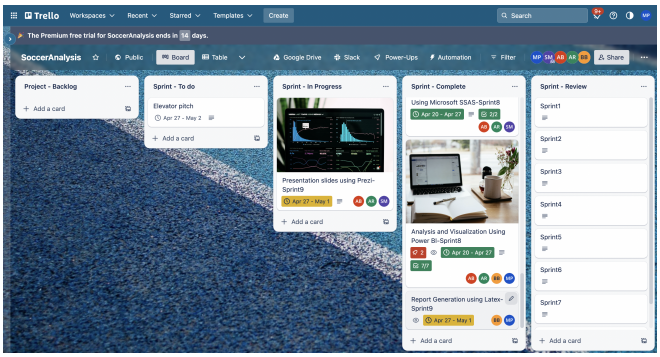


Fig. 9. Trello Dashboard

Trello Link:https://trello.com/b/dlrWxVA7/socceranalysis/

2. **Minutes of Meeting**

Meeting minutes are an essential part of any project, providing a written record of the discussions, decisions, and action items from each meeting. In the case of our project, we held regular meetings using Zoom, with each sprint consisting of 1- 2 calls. Throughout the project, we maintained a document titled ”Project Meeting Minutes,” where we recorded the minutes of each meeting. This document allowed us to keep track of the progress made, the challenges faced, and the decisions taken at each meeting. A detailed record of our meetings ensured that everyone was on the same page and that any action items were promptly addressed. Overall, the meeting minutes were essential in helping us stay organized and focused throughout the project.

VII. Version Control

We were able to leverage the version control system, ie, **GitHub**, to work on the ETL pipelines efficiently. By using separate branches for each task, team members could work independently without interfering with each other's code. Each branch contained the changes related to a particular task, and team members could review each other's changes using pull requests before merging them into the main branch. This ensured the code was thoroughly reviewed and issues were caught early on. Furthermore, any conflicts during the merging process were resolved quickly and efficiently.

Pair Programming was possible with the help of Github. Multiple team members could work on the same code base simultaneously, providing real-time feedback and assistance to each other. Using version control in this way helped the team split tasks more efficiently and complete the ETL pipelines in a timely manner. It also helped ensure that changes were thoroughly reviewed and the codebase remained stable and organized. Thus, leveraging the features of GitHub helped the team to collaborate effectively and produce high-quality ETL pipelines.

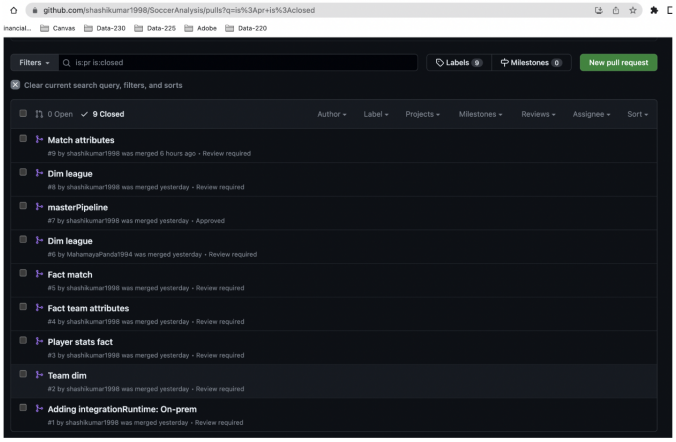


Fig. 10. Pair Programming in GitHub

GitHub Link:https://github.com/shashikumar1998/SoccerAnalysis/

VIII. Data Analysis and Visualization

*1) **Data analysis using SSAS**:* Two kinds of dimensional models are explored:

(i) **SSAS Tabular model:** The tabular model represents an in-memory data view and gives performance benefits by compressing large data sets. Since we can store the data in memory, it provides a performance advantage compared to reading the data from a database.

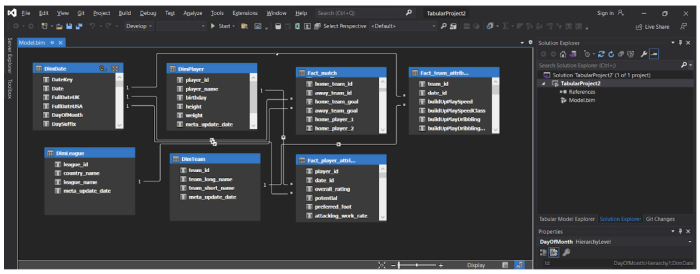


Fig. 11. SSAS Tabular model

(ii) **SSAS Multidimensional model (Data Cubes):** The data is imported, and dimensional modeling is done where the measures and dimensions are chosen. Later, the cube is deployed in the SQL server's analysis server to perform query and aggregation of attributes.

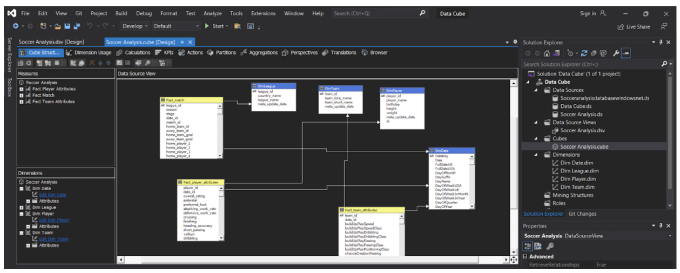
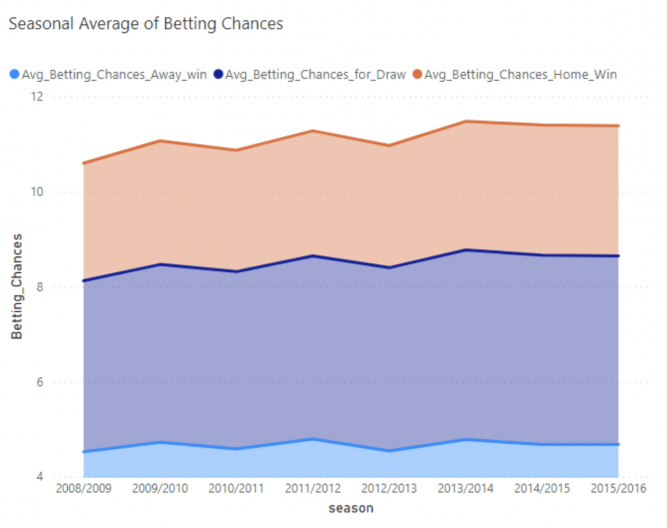


Fig. 12. SSAS Multidimensional Model

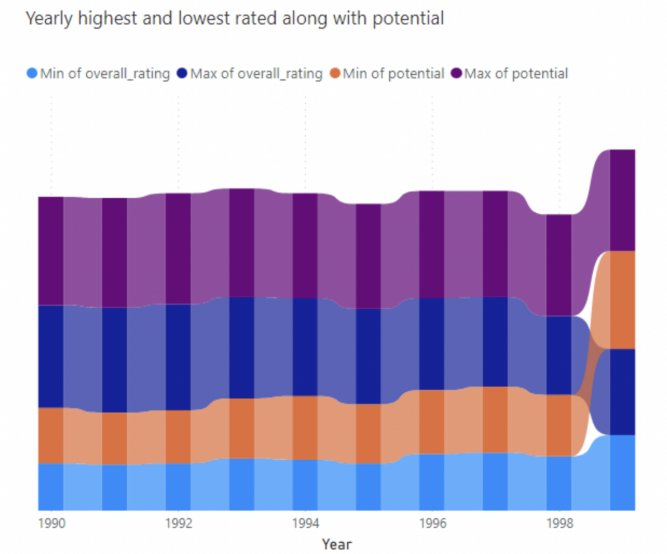
The analysis results are presented below in a suitable format using a data visualization tool, i.e., Power BI.

*2) **Data Visualization**:*

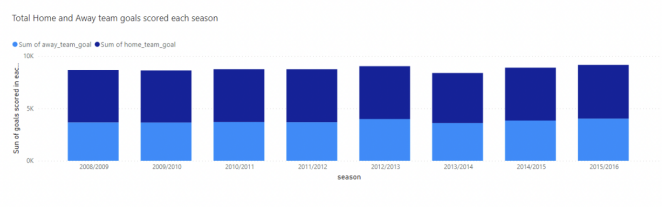
1. The first visualization states the chances of the home team winning, the team draw, or the away team winning for nine seasons. This can be utilized to understand how home and away teams play with each other. This is just a preliminary examination of existing analytical data before performing an in-detail analysis with information.



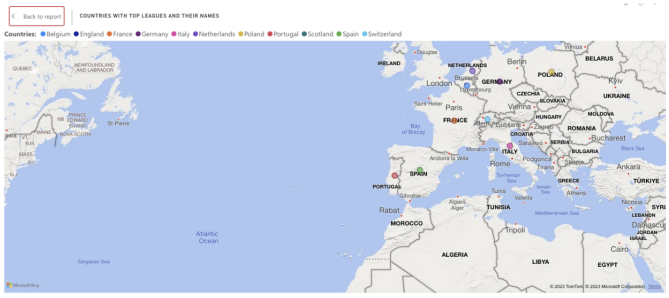
2. The visualization below provides us with data on the players who are part of professional football each year with the highest and lowest ratings and potential. This helps us understand and evaluate the productive year and the rise of the quality of players. Along the line, we also can find if there is



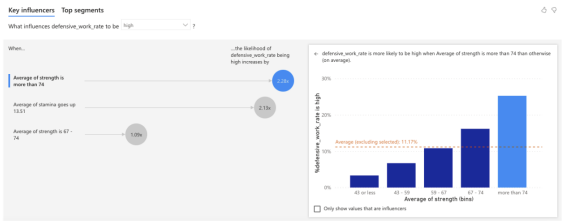
3. The following visualization provides us with the detailed goals scored by home and away teams each season, again evaluated for nine seasons. This helps us understand if the season provided the home or away team an attacking edge. We have seen coaches complaining about the dampness of the away team pitches providing aid in the defensive structure of away teams.



4. The world map below shows which countries are part of the FIFA and UEFA competitions, and they also are to be part of the top 10 leagues in the world. The legend says about countries' names with color coding, and corresponding points are pointed over the map; hovering over it gives us details about league names called in that nation.

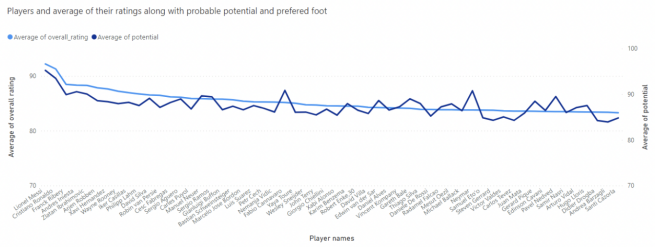


5. The following graph is a feature in Power BI that allows us to understand the key influencers of the information we require. In the following, we implemented a very important defensive work rate feature, which decides how a team can hold other teams from scoring. If a manager needs to pick a player who supports him in building the defense, the qualities mentioned below influence the most. Through the detailing, the manager can pick the player with suitable attributes.

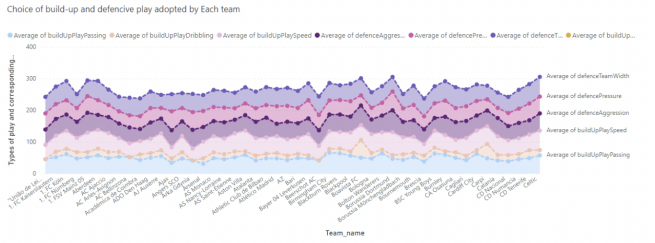


6. Similar to the above, the key influencer graph below shows the attacking attributes of a player and what makes him a threat in the team's attack. A feature also allows users to change the attacking work rate from high to low; each has its benefits. For example, if a manager requires an attacking player who tracks back or joins midfield, his work rate should be high, or if the player needs to be only forward and at the edge of the opponent's box, then the work rate can be low.

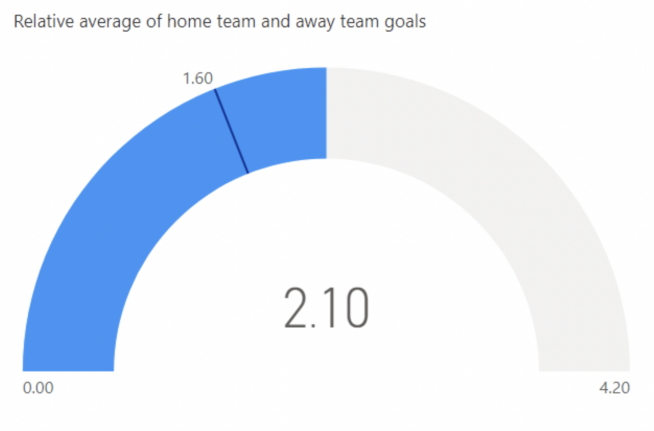
7. The following line graph shows two different lines: one is the average rating, and the other tends to be the potential of a player. This tracks the player ratings and their potential, aids in quickly knowing about the player's position in world rankings overrating and what can be his potential. We can also analyze the player's age and the possibility of renewing the contract. If the potential falls below the rating, which means the player is nearing the end of a professional football career, or if the player's potential has skyrocketed from his rating, then there is a chance of wonder kid brewing, etc.



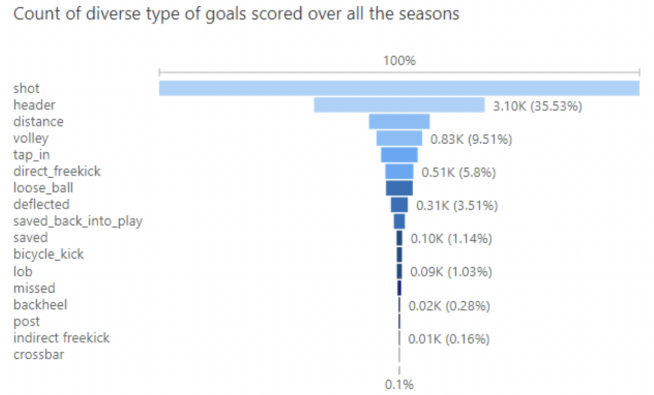
8. The below complex graph is an area chart (stacked), which shows each team's adjustment to a different style of play and their choice of play in both build-up and defense. For example, the defense has three other plays. Similarly, build-up play has three diverse sports. The average is from the information we hold of each style of play and the overall average evaluation of that value. For example, we can know which defense tactic works best if a team adopts speed build-up play. This type of analysis can be drawn from the below.



9. The below goal meter shows how many goals are being scored by the home and away teams on average in a match over a seasonal period. If the average of home team goals is lower than away team goals, then the away visits are becoming facile and dominant. Still, from the below case, we can see that the average of home team goals is higher than away teams, which means the home domination is still in play, and no away domination establishment is yet to be seen.



10. The below visualization describes the type of diverse goals score. The percentage seen is in comparison with the first type of goal. Next, we find that the shots from forward are the highest, and next comes the headers, followed by distance shots which mean outside-the-box shots. Finally, the crossbar and then-in are the least scored of all. This helps in understanding how many types of shots depend on a goal scorer and how many are left to the brilliance of the scorer and the rest to luck.



IX. Conclusion

The project aimed to analyze the soccer team's performance through data-driven technologies. Various data loading techniques, such as Bulk insert and Upsert using Azure Data Factory, were used to analyze data in the system. This allowed the team to use advanced analytics techniques to identify patterns that informed team strategy and improved performance. Hence, the project demonstrated the effectiveness of data-driven technologies in improving soccer team performance.

X. Lessons Learned

1. We got hands-on experience in dealing with the whole data lifecycle.

2. Worked with tools taught in class like MySQL and explored new tools like CosmosDB for Mongo, Azure Data

Factory, Azure Dedicated SQL Pools, Power BI, etc., to handle data and perform analysis.

3. We practiced data modeling and dimensional modeling to create a new data project.

4. Agile development methodology can be used to develop data-related projects, and tools such as Trello can help with project management.

5. New presentation and documentation tools like Prezi and Latex were explored and used for this project.

6. Gained experience in pair programming, version control, project management, and teamwork.

XI. Technical Difficulty

1. We are implementing the concepts taught in this course using various on-prem and cloud tools like Azure Data Factory, CosmosDB, Azure dedicated SQL pools, etc., which are new for our team and were a steep learning curve in the beginning.

2. One of our data sources was MySQL which was implemented on-prem. For accessing this database from Azure Data Factory for ETL, we had to install a self-hosted integration runtime on the local machine to act as a gateway to the Azure cloud.

3. Accessing the data warehouse, which was on the cloud, was a challenge because Azure, by default, doesn't provide public access to its resources for which we had to individually add IP addresses of each machine to make sure everyone on our team got access to perform analysis of the data.

4. Deploying the SSAS multidimensional cube to the Azure Analysis service was impossible, so we had to deploy it to the local instance of the SQL server.

5. Power BI is a useful tool for visualizing and is powerful in performance, but the issue arises with licensing the same tool. There are very restricted features with this tool in a free trial. To share the visualizations and reports with the entire team, we were forced to begin a free trial for 60 days.

XII. Novelty

In choosing the dataset and concepts for the project, we sought to balance novelty and practicality. As a result, we showcased our ability to apply course concepts while exploring new technologies relevant to the project. Furthermore, our project analyzed the various factors contributing to a team's success, like 'how home vs. away matches affect the outcome', 'player stats like goal impact metrics', etc.

XIII. Impact

The project impacts managing companies, team players, wager candidates, or fanatics to make data-driven decisions from the insights obtained. It helps to make high-quality decisions and yield better results. Managing companies must make informed decisions while selecting players and consider many factors. Deriving insights from comprehensive analysis saves them the necessary resources that could go in vain. Players can have better diet plans and training strategies depending on the strength and weaknesses of the opposition. Wager candidates can have safer bets and possess higher chances of winning.

XIV. Future Scope

1. Incorporating more real-time data to improve our data analysis and reporting accuracy and relevance.

2. Implementing AI/Machine Learning algorithms to gain deeper insights from the data and identify patterns that can help predict game outcomes.

3. Including social media data and sentiment analysis to better understand fan engagement and its impact on team performance and to create more personalized experiences for users.

XV. Grammarly and Plagiarism Checks

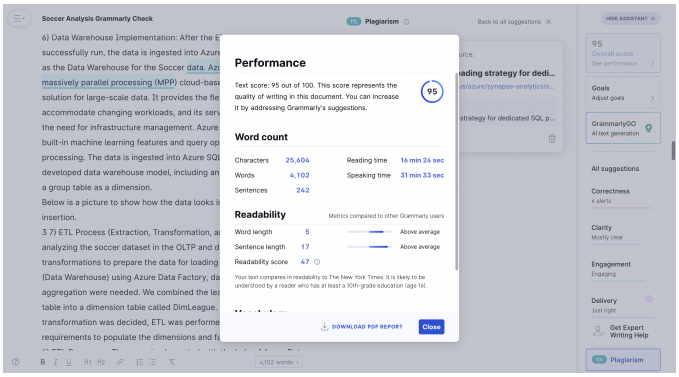


Fig. 13. Grammarly Check

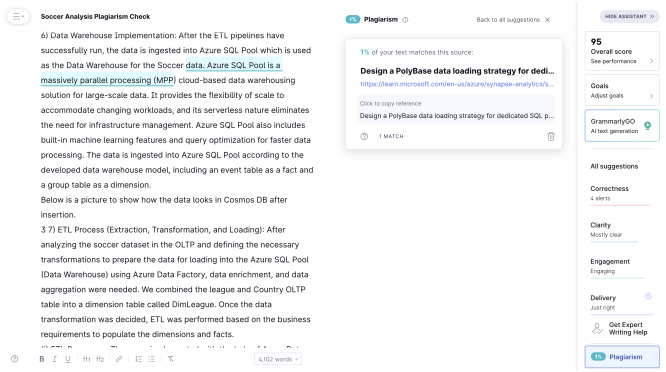


Fig. 14. Plagiarism Check

References

[1] *https://www.frontiersin.org/articles/10.3389/fpsyg.2018.02416/full*
[2] *https://www.linkedin.com/pulse/data-science-football-how-data-driven-approach-has-changed-joshi/*
[3] *https://arxiv.org/abs/2102.07545*
[4] *https://medium.com/trends-in-data-science/squad-selection-and-match-analysis-in-soccer-using-big-data-62c21ba9b470*
[5] *https://ieeexplore.ieee.org/document/10025111/metricsmetrics*

XVI. Team Members / Roles

Team Members	Roles
Anitha Balachandran	Data gathering, data cleansing, data analysis
Aradhya Alva Rathnakar	Data gathering,data modeling, reporting
Bhavan Kumar Basavaraju	Data gathering, data cleansing, ETL
Mahamaya Panda	Data gathering, data modeling, data analysis
Shashi Kumar Kadari Mallikarjuna	Data gathering, ETL, data analysis, reporting

XVII. Appendix

Criteria	Comments
Version Control Use of Git / GitHub or equivalent; must be publicly accessible	Git/Github was used for version control where we linked the Azure data factory pipelines so that multiple team members could work on creating pipelines for different dimensions and fact tables and eventually merged data to the main branch.
Significance to the real world	Football is a sport that in some places is treated as a religion and even as a lifestyle. From the day the sport was born, the world has been evolving and technology has taken over most of the themes. However, we propose the use of technology to make the sport more fun and lively than ever. Our work here determines the way how analysis can reduce the hassle of long documents and maintaining records.
Lessons learned Included in the report and presentation? How substantial and unique are they?	We learned the usage of various new tools and cloud technologies for better handling of data and performing analysis. We got a chance to practice different concepts taught in class like data modeling, dimensional modeling, etc. The lessons we learned are mentioned in much detail in this report.
Innovation	As we are aware in this current world technology overtaking almost every industry, and innovation is always possible in every field. In football, we have seen football analysis as a key aspect that was performed in support of team analysis, pitch analysis, etc. We through this project try to create a common layman platform that can be utilized by any FIFA nation or even a football fan. The innovation in this factory includes the usage of diverse databases to procure swift responses, different visual representations focusing from player to league level information, and other significant improvements to the existing systems,
Teamwork	The work was split among the team members equally and everyone got a chance to work on all the tasks to get an exposure in terms of using different tools.
Technical difficulty	We faced issues in terms of integrating data from various sources, where we had to download a self-hosted integrated runtime to access data from the local system, and faced network-related issues when connecting to the data warehouse for which IP address had to be added explicitly. Deploying SSAS multidimensional cube to Azure analysis service was not possible for which we had to deploy it on SQL server on our local machine.
Practiced pair programming?	Using GitHub, we were able to practice pair programming where if one of our team members was stuck with some implementation, we could work together on that branch. Once a pipeline is ready for deployment, we create pull

	requests for other team members to review before merging the changes to the main branch to make sure it is working as intended by testing the functionality.
Practiced agile / scrum (1-week sprints)? Submit evidence on Canvas - meeting minutes, and other artifacts	We split the project into multiple story points and developed it in Agile sprints. Each sprint cycle was 1 week long and the tasks were split between team members. Trello was used to track the user stories and sprints
Used Grammarly / other tools for language.	Grammarly was used to check the verbosity and the grammatical errors in our report. Along with that, the google Docs has a very good built-in feature to point out a suggestion for any wrong usage of words or auto-correcting the wrong wordings.
Slides	Project presentation slides were created using Prezi with all the essential content and are made visually appealing.
Report Format, completeness, language, plagiarism,	The report follows IEEE format and includes all the details about the project
Used unique tools	Prezi was used to create the project presentation slides and Latex is being used to create the report.
Performed substantial analysis using database techniques The project must include an analytics component	Using the data loaded into the data warehouse, we used SQL queries to get valuable insights into the data and displayed the insights using PowerBI
Used a new database or data warehouse tool not covered in the HW or class	Azure dedicated SQL pool is used as a data warehouse using which analysis was performed
Used appropriate data modeling techniques	Understood the business requirements, created ER models, and normalization was performed to make sure all the tables were in 3NF before loading the data into the OLTP databases. Performed dimensional modeling to identify dimensions and fact table
Used ETL tool	Azure data factory (ADF) was used to extract the data from the data sources (CosmosDB and MySQL), transform the data, and load it into Azure's dedicated SQL pool. ADF was also used to schedule nightly runs to load the data from the OLTP databases to the data warehouse
Demonstrated how Analytics support business decisions	Using data from Azure's dedicated SQL pool(data warehouse), we have written queries, built SSAS multidimensional cube, and visualized that data using powerBI which is a visualization tool.
Used RDBMS	MySQL, which is an RDBMS, was used as one of our data sources which has the transactional level of data stored like the match details, team stats, etc.
Used Data Warehouse	Azure dedicated SQL pool, which is an analytical

	database offered by Azure was used as the data warehouse in which the processing power could be set dynamically based on our usage using DWUs(Data warehousing units)
Includes DB Connectivity / API calls	MySQL ODBC driver was used to connect to MYSQL from Python to load the CSV data
Used NoSQL	Azure CosmosDB for MongoDB has been used to store the players' information which is used as our master data management of players for analysis.