JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 4 No 1 June 2025
Hal: 1-9
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v4i1.284

A Data-Driven Approach to Business Intelligence for Evaluating Football Player Performance

Lam Mai[1,*], Thu-Thuy Tran[2], Duc-Hien Nguyen[3]

[1,2,3] The University of Danang, Vietnam - Korea University of Information and Communication Technology, Danang, Vietnam
Email: [1,*]mlam@vku.udn.vn, [2]ndhien@vku.udn.vn
(*Email Corresponding Author: mlam@vku.udn.vn)
Received: 18 December 2025 | Revision: 14 February 2025 | Accepted: 19 May 2025

Abstract

Analyzing football players' performance is a crucial focus in modern sports science, providing insights into player efficiency and team strategies. This paper proposes a comprehensive framework for evaluating player performance by integrating statistical metrics, match data, and advanced analytics techniques. Key performance indicators (KPIs), including passing accuracy, goal contributions, and defensive actions, are analyzed alongside contextual factors such as match conditions and opposition strength. Using a dataset of per-90-minute statistics for the 2022-2023 season, this study covers players from top European leagues: the Premier League, Ligue 1, Bundesliga, Serie A, and La Liga. The proposed model offers coaches, analysts, and researchers actionable insights to enhance player development and optimize team strategies.

Keywords: Data warehouse, Football players' performance, ETL, OLAP, Multidimensional Data

1. INTRODUCTION

Football, often referred to as the "beautiful game," is not only the world's most popular sport but also a dynamic field for scientific research and technological innovation. Understanding player performance is vital for optimizing individual contributions and achieving collective team success. Traditional methods of performance evaluation, such as subjective observation, have given way to data-driven approaches that leverage advancements in sports analytics and machine learning.

The analysis of football player performance encompasses a wide array of factors, including technical skills, physical attributes, and decision-making under pressure. With the advent of wearable technology, video analysis software, and comprehensive data collection systems, it has become possible to capture granular details about players' actions during matches. However, translating this wealth of data into actionable insights remains a significant challenge.

A database can be defined in various ways, but fundamentally, it is a collection of data designed to meet specific business requirements, with each piece of data having logical interconnections [3]. As [4] explains further, a database is a structured collection of data that can be centrally managed, integrated, and stored. From these definitions, it can be inferred that a database is a logical grouping of interconnected data stored systematically to enable its transformation into meaningful information for organizations. The processes of management and control differentiate raw data from databases, with the latter providing a structured approach to organizing data sets.

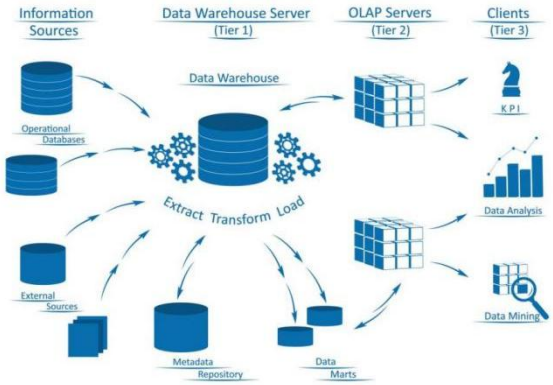

Figure 1. Business Intelligence Architecture Diagram

According to [5], ETL tools are critical for migrating data between databases, creating data marts and data warehouses, and converting data formats or types. As outlined by [6], the ETL process consists of three main phases: Extraction: This is the initial stage where data is collected from internal and external sources. At this point, logical differences might emerge, particularly when historical data is introduced into an empty data warehouse. This phase ensures that the data warehouse is updated with new data, aligning it with business intelligence requirements and decision support systems; Transformation: This phase focuses on cleansing and transforming the data to improve its quality. It resolves inconsistencies and enhances the accuracy of data collected from various sources; Loading: The final phase

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 4 No 1 June 2025
Hal: 1-9
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v4i1.284

involves loading the processed data into data warehouse tables, making it ready for analysis and generating business intelligence reports.

As [6] further explains, ETL tools, or data integration applications, facilitate these processes by extracting, transforming, and preparing data for warehouse loading. These tools also support decision-making systems by enabling professionals to analyze and visualize data. Among ETL tools, Pentaho is particularly effective due to its intuitive user interface and fast batch processing, making it suitable for constructing data warehouses from raw data [7]. Additionally, Pentaho is compatible with the Kimball methodology, which enhances its utility for ETL processes [8].

As defined by [9], "A data warehouse is a subject-oriented, integrated, time-variant, and non-volatile collection of data that supports management's decision-making processes." According to [10], this definition can be broken down into the following aspects: Subject-oriented: Data is organized based on business functions; Integrated: Data from multiple sources is combined for unified analysis; Time-variant: The warehouse stores historical data to track past transactions; Non-volatile: The data is read-only, ensuring that it does not modify original records.

In essence, a data warehouse is a centralized database that consolidates business data for analysis. One of the primary challenges of creating a data warehouse lies in collecting and organizing massive amounts of data from various systems to support business intelligence. The ETL process is fundamental to this task, enabling efficient data loading into the warehouse [11]. This paper addresses this challenge by proposing a systematic methodology for performance analysis that integrates quantitative and qualitative dimensions. By focusing on KPIs tailored to specific player roles and situational contexts, the study seeks to provide a deeper understanding of what constitutes effective performance.

Our contributions focus on:

- Developing a comprehensive framework for analyzing football player performance through the use of advanced analytics tools (SSIS, SSAS, SSRS).
- Integrating contextual variables and key performance indicators (KPIs) to construct a holistic model for performance assessment.
- Demonstrating the application of this framework using publicly available datasets from Kaggle .

The findings are intended to assist coaches and analysts in making evidence-based decisions to enhance player development, optimize tactics, and gain a competitive edge in modern football. The insights derived from this research aim to provide actionable strategies for enhancing player performance, optimizing team tactics, and fostering sustained competitiveness in football. The remainder of the paper is organized as follows: Section II introduces the proposed methodology, detailing the integration of key performance indicators (KPIs), contextual variables, and advanced analytics techniques. Section III presents the experimental results, showcasing the application and validation of the framework. Finally, Section IV concludes the study by summarizing key findings and discussing potential directions for future research.

Figure 2. Our star schema with one fact and eight dimensions

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 4 No 1 June 2025
Hal: 1-9
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v4i1.284

2. RESEARCH METHODOLOGY

2.1 Data Collection

Match statistics and player data are sourced from Kaggle's extensive football datasets [13], which offer a rich repository of over 25,000 rows and 124 columns. These datasets comprehensively document player actions, match outcomes, and contextual elements such as weather conditions and opposition strength. They serve as a detailed source for analyzing player performance and team strategies.
The dataset's structure includes 124 columns that provide granular insights into various aspects of football matches. Some of the key columns are as follows:

Table 1. Types of database

ID	Attribute	Description
1	Rk	Represents the rank of the player.
2	Player	The name of the player.
3	Nation	Indicates the player's nationality.
4	Pos	The position in which the player primarily plays (e.g., Forward, Midfielder).
5	Squad	The name of the squad the player represents.
6	Comp	The league in which the player's squad competes.
7	Age	The player's current age.
8	Born	The year the player was born.
9	MP	Matches played by the player.
10	Starts	The number of matches in which the player was in the starting lineup.
11	Min	Total minutes played by the player.
12	90s	A calculation of minutes played divided by 90, representing full games played.
13	Goals	The number of goals scored or allowed by the player.
14	Shots	The total number of shots attempted by the player (excluding penalty kicks).

The dataset also includes advanced metrics such as passing accuracy, assists, xG, xA, and defensive actions, allowing for an in-depth analysis of player and team performance. This expansive structure supports various analytical workflows, including predictive modeling and performance benchmarking.
To enhance the dataset, additional sources such as publicly available APIs and repositories are integrated. These supplementary resources enrich the dataset by adding dimensions like weather conditions, match attendance, and opposition strength, enabling a holistic view of the factors influencing match outcomes. Together, the data forms a robust foundation for statistical analysis and strategic insights in football analytics.

2.2 Data Integration

Subtitle 2 can also contain the method of solving the problem, as well as the stages of the method. In the manuscript, quotation numbers are sequentially in square brackets [3], as well as tables of numbers and numbers sequentially as shown in table 1 and figure 1.

SQL Server Integration Services (SSIS) is employed to extract, transform, and load (ETL) data from diverse sources. The ETL process ensures data quality, consistency, and completeness by cleaning raw data, resolving discrepancies, and standardizing formats. The integrated dataset is stored in a relational database for further analysis. Figure 3 shows the ETL control flow, means show the sequence of intergrating data using SSIS. The ETL process is shown on Figure 3 – 12.

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 4 No 1 June 2025
Hal: 1-9
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v4i1.284

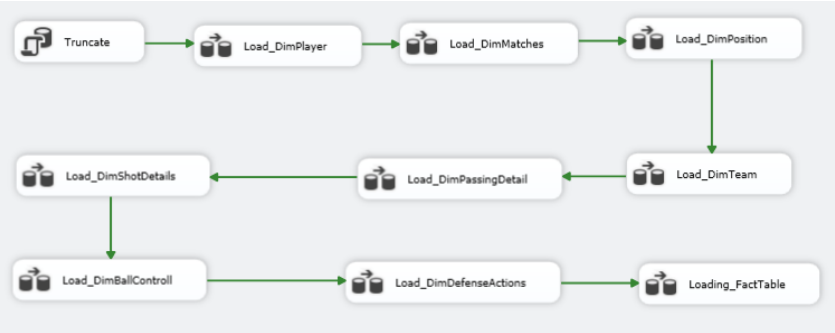

Figure 1. ETL control flow

Figure 4 show Sql command for Truncate, this is the first step guarantee that all the incomming data will not conflict with old data.

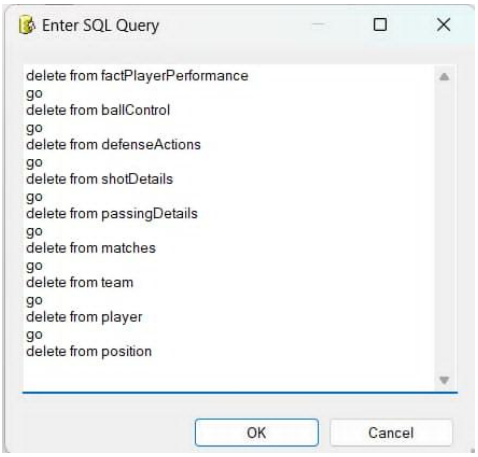

Figure 2. Sql command for Truncate

In Figure 5, we create Players dimension to store some informations about Players like name, nation, born, age,… from original data source.

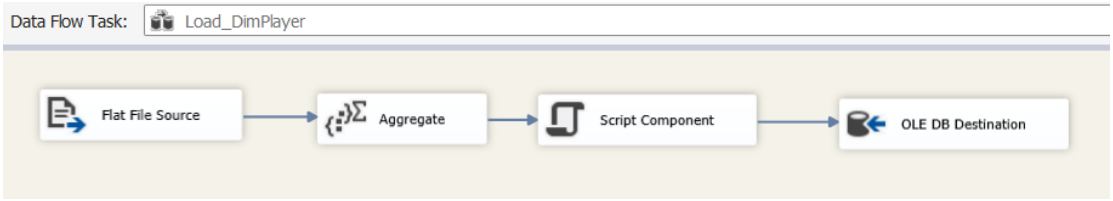

Figure 3. Players Dimension ETL

In Figure 6, we create Matches dimension to store some informations about Match like Matches played(mp), Matches started(starts), Minutes played(min),… from original data source.

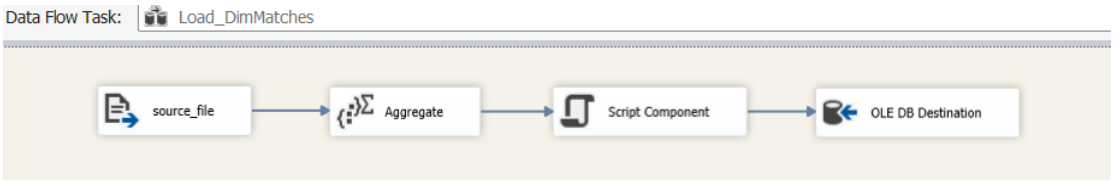

Figure 4. Matches Dimension ETL

In Figure 7, we create Position dimension to store the Position's name(pos) from original data source.

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 4 No 1 June 2025
Hal: 1-9
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v4i1.284

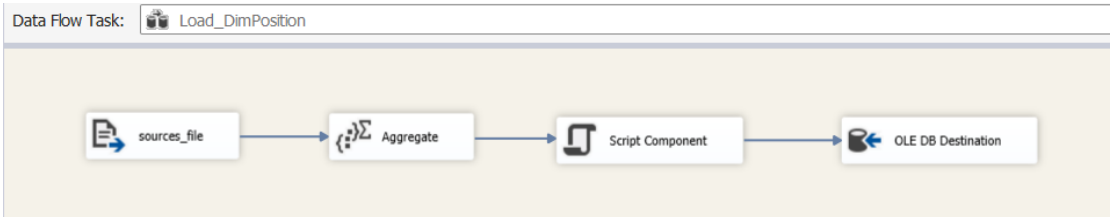

Figure 5. Position Dimension ETL

In Figure 8, we create Team dimension to store the name of team which player are playing for. In Figure 9, we create Passing Detail dimension to store some informations related to Passing like Passes attempted(PasAtt), Live-ball passes (PasLive), Dead-ball passes(PasDead), Passes completed(PasTotCmp), Passes attempted(PasTotAtt),Pass completion percentage (PasTotCmpPercentage) from original data source.

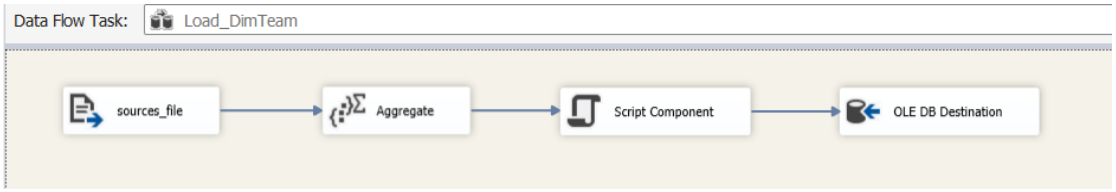

Figure 6. Team Dimension ETL

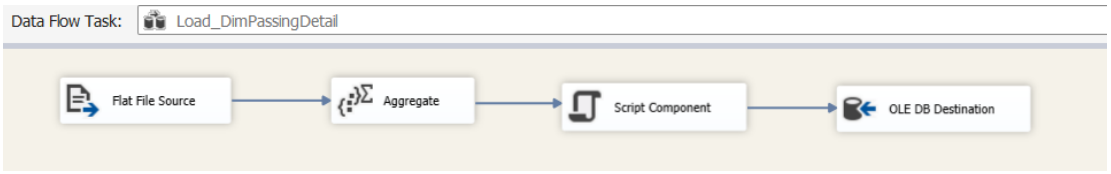

Figure 7. Passing_Detail Dimension ETL

In Figure 10, we create Shots Detail dimension to store some informations related to Shots like Shots total(shots), Shots on target(soT), goals per shot(GperSh),… from original data source.

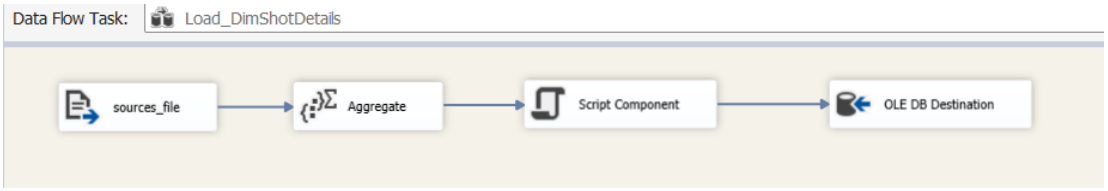

Figure 8. Shot_Detail Dimension ETL

In Figure 11, we create Ball Controll dimension to store some informations related to Ball Controll from original data source.

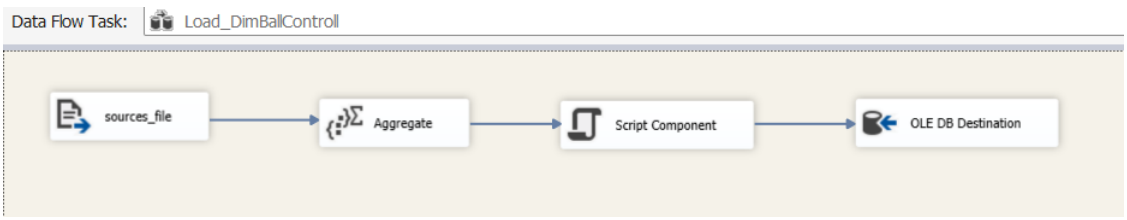

Figure 9. Ball_Controll Dimension ETL

In Figure 12, we create Players dimension to store some informations about Players like number of times blocking the ball by standing in its path(Blocks), Number of times blocking a shot by standing in its path(BlkSh), Number of times blocking a pass by standing in its path(BlkPass),… from original data source.

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 4 No 1 June 2025
Hal: 1-9
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v4i1.284

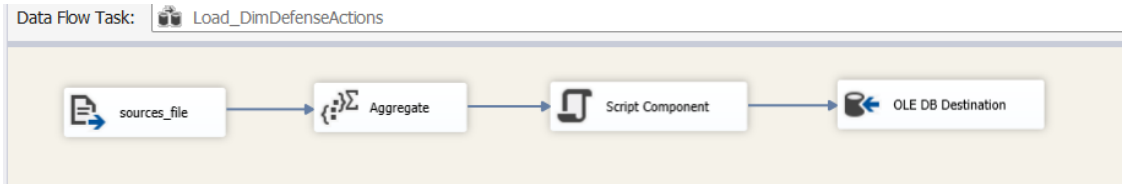

Figure 10. Defense_Actions Dimension ETL

In Figure 13, we create a Fact table to store key performance metrics related to players' actions during matches. This table is designed to capture quantitative data extracted from the original data source, enabling detailed analysis and reporting of player performance.

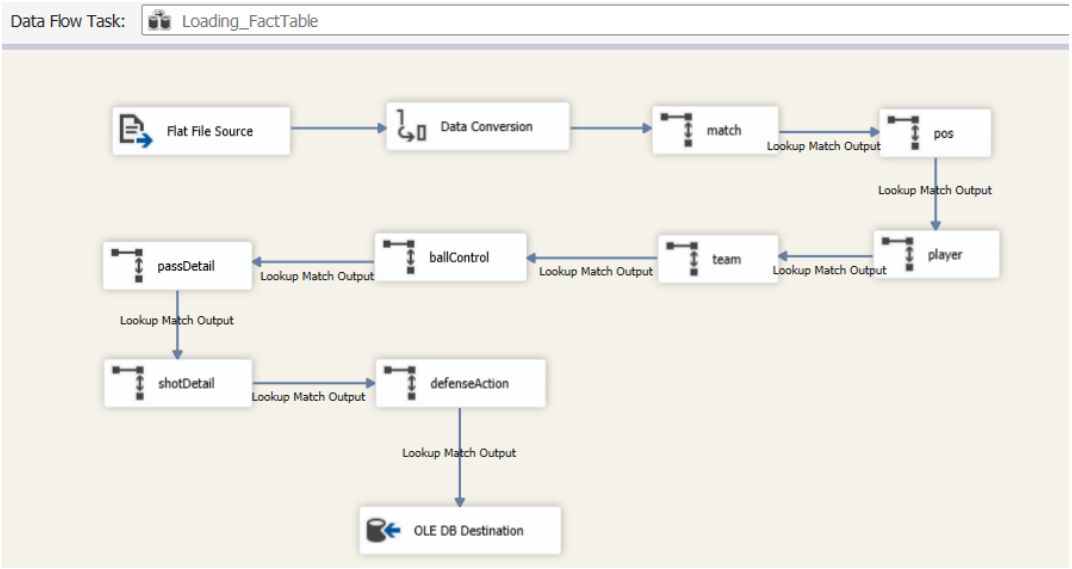

Figure 11. The sequence of ETL process

2.3. Data Analysis

SQL Server Analysis Services (SSAS) is utilized to perform multidimensional analysis, enabling the exploration of complex relationships between performance metrics and contextual factors. Key Performance Indicators (KPIs), such as passing accuracy, defensive contributions, and goal involvement, are analyzed using OLAP cubes to identify patterns and trends.

3. RESULTS AND DISCUSSION

Once the ETL process is complete, we can utilize the data warehouse to create reports. In this case, Power BI tools will be used. After establishing a connection to the data warehouse, we can organize the data into a star schema, as illustrated in Figure 2. With the star schema in place, we can proceed to generate reports, as depicted in Figures 14 through 18.

As shown in figure 14, this dashboard provides a comprehensive analysis of football player performance across various metrics, offering valuable insights for team management and strategy development. The disciplinary analysis identifies players with the highest number of yellow and red cards, highlighting areas for improvement in maintaining discipline. Goal-scoring metrics reveal top performers and provide a clear view of team efficiency in offensive play, while the age distribution showcases a balance of experienced veterans and young talent within the squads. With a total of 2,613 players and 98 teams analyzed, the dashboard provides a robust dataset for benchmarking individual and team performances. By comparing cumulative metrics such as goals, cards, and ages, this analysis supports data-driven decisions to enhance overall team performance and identify key areas of focus for future improvement.

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 4 No 1 June 2025
Hal: 1-9
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v4i1.284

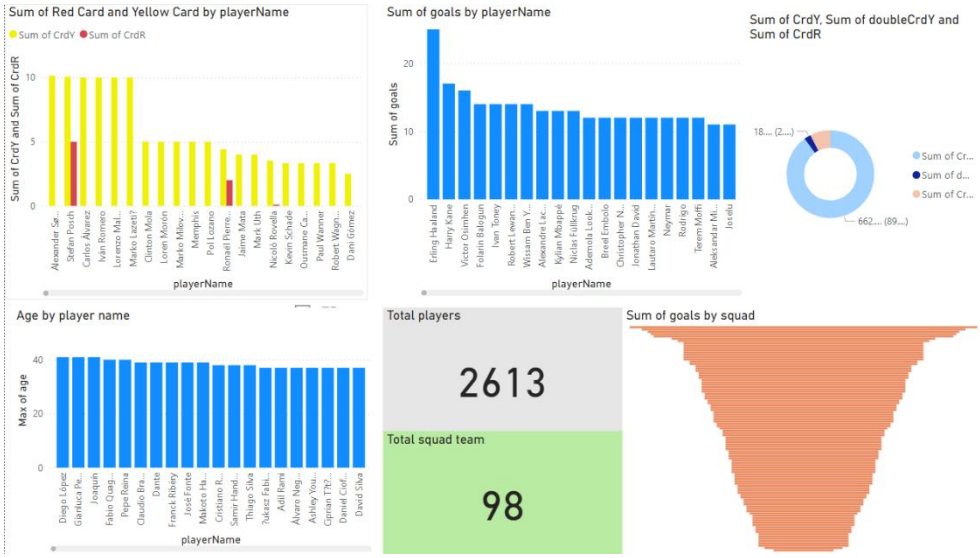

Figure 14. Dashboard show general information of project

Figure 15 provides a clearer description of what the chart represents, highlighting the comparison between the two types of fouls for individual players.

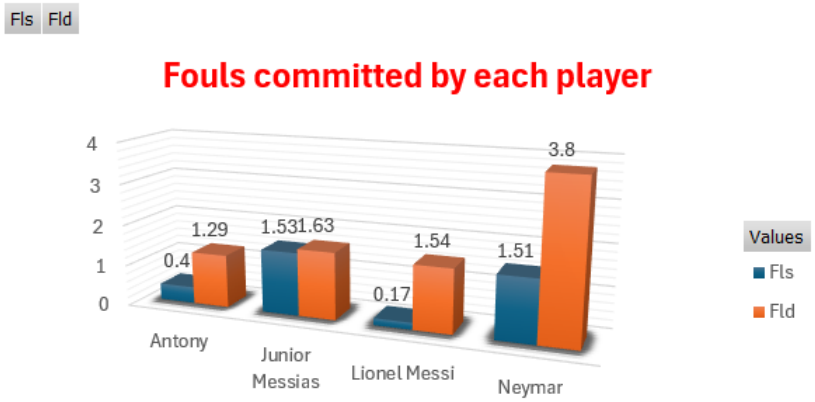

Figure 15. Comparison of Fouls Suffered (Fls) and Fouls Committed (Fld) by Each Player

As shown in figure 16, the chart provides more context by emphasizing that the chart analyzes the disciplinary actions taken against players in the form of yellow and red cards due to violations.

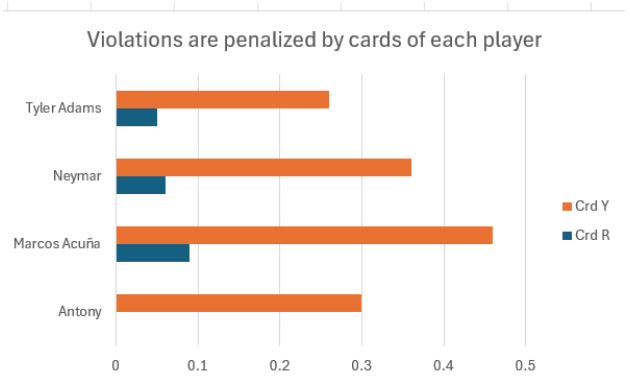

Figure 16. Distribution of Yellow (Crd Y) and Red Cards (Crd R) Issued to Each Player

In Figure 17, the chart show the performance contributions of several players by comparing three key metrics: goals scored, total touches, and assists provided. Each player is represented with distinct bars corresponding to these metrics, allowing for a detailed evaluation of their overall impact on the game. The blue bars indicate the number of goals scored.

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 4 No 1 June 2025
Hal: 1-9
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v4i1.284

The orange bars represent the total number of touches. Lastly, the green bars denote the number of assists, showcasing their role in creating scoring opportunities for teammates. This comparative analysis provides a comprehensive overview of individual performances across these crucial aspects of the game.

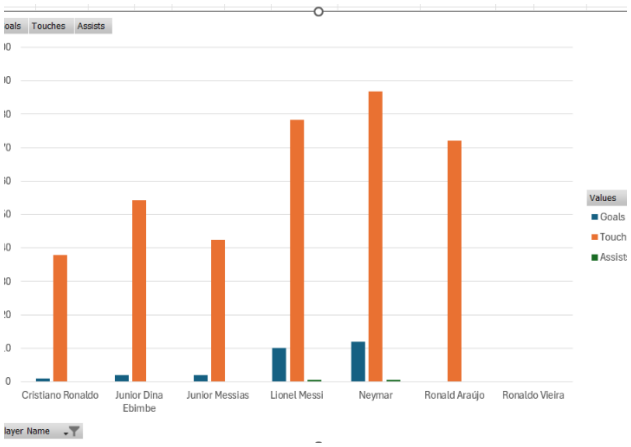

Figure 17. Player Contributions: Goals, Touches, and Assists

In Figure 18, bar chart displays the total number of yellow (CrdY) and red (CrdR) cards received by various players. Yellow bars represent yellow cards, and red bars represent red cards. Notable players like Alexander Sørloth lead with the highest counts.

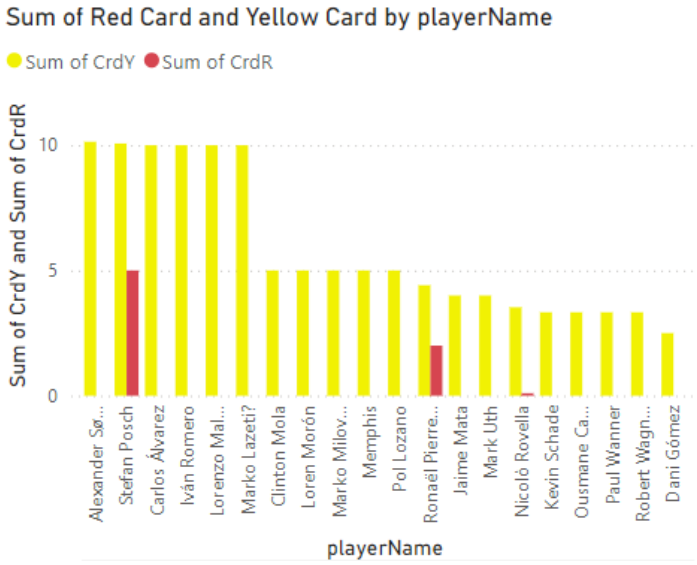

Figure 18. Player Card Statistics Visualization

4. CONCLUSION

This study highlights the critical role of data warehousing in evaluating football player performance through a data-driven approach to business intelligence. By integrating ETL processes with tools like SSIS, SSRS, and Power BI, organizations can efficiently analyze data and create insightful visualizations. Power BI, in particular, enhances reporting through interactive dashboards and advanced analytics, making it invaluable for performance evaluation. These methods, when extended to big data technologies, offer even greater potential for analytics and strategic decision-making in sports and beyond.

REFERENCES

[1]. F. Almeida, "Concepts and fundamentals of data warehousing and OLAP," ISSUU Publishing, 2017. [Online]. Available:

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 4 No 1 June 2025
Hal: 1-9
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v4i1.284

https://www.researchgate.net/publication/319852408_Concepts_and_Fundaments_of_Data_Warehousing_and_OLAP.

[2]. N. Sharma, A. Iyer, R. Bhattacharya, N. Modi, and W. Crivelini, "Getting started with data warehousing (Draft)," IBM Corporation, 2012. [Online]. Available: https://freecomputerbooks.com/Getting-Started-With-Data-Warehousing.html.

[3]. T. M. Connolly and C. E. Begg, Database systems: a practical approach to design, implementation, and management, 6th ed., global ed. Boston, MA: Pearson, 2015.

[4]. J. W. Satzinger, Systems analysis and design in a changing world, 7th ed. Boston, MA: Cengage Learning, 2015.

[5]. R. Kimball, Ed., The data warehouse lifecycle toolkit, 2nd ed. Indianapolis, IN: Wiley Pub., 2008.

[6]. C. Vercellis, Business intelligence: data mining and optimization for decision making. Chichester, U.K.: Wiley, 2009.

[7]. D. Stooder, TDWI Best Practices: Improving Data Preparation for Business Analytics. Hitachi Vantara. [Online]. Available: https://www.hitachivantara.com/enus/pdf/analyst-content/improving-data-preparation-forbusiness-analytics-tdwi-best-practices-report.pdf.

[8]. R. Kimball and J. Caserta, The data warehouse ETL toolkit: practical techniques for extracting, cleaning, conforming, and delivering data. Indianapolis, IN: Wiley, 2004.

[9]. W. H. Inmon, Building the data warehouse, 4th ed. Indianapolis, IN: Wiley, 2005.

[10]. L. Yessad and A. Labiod, "Comparative study of data warehouses modeling approaches: Inmon, Kimball and Data Vault," in 2016 International Conference on System Reliability and Science (ICSRS), 2016, doi: 10.1109/ICSRS.2016.7815845.

[11]. S. Vyas and P. Vaishnav, "A comparative study of various ETL process and their testing techniques in data warehouse," Journal of Statistics and Management Systems, vol. 20, no. 4, pp. 753–763, Jul. 2017, doi: 10.1080/09720510.2017.1395194.

[12]. R. Kimball and M. Ross, The data warehouse toolkit: the definitive guide to dimensional modeling, 3rd ed. Indianapolis, IN: John Wiley & Sons, Inc., 2013.

[13]. "Football Player Stats 2022–2023," Kaggle. [Online]. Available: https://www.kaggle.com/datasets/vivovinco/20222023-football-player-stats.

[14]. L. Mai, X.-Z. Chen, C.-W. Yu and Y.-L. Chen, "Multi-view vehicle re-identification method based on Siamese convolutional neural network structure", Proc. IEEE Int. Conf. Consum. Electron., pp. 1-2, 2020.

[15]. L. Mai, X.-Z. Chen and Y.-L. Chen, "Multi-oriented license plate detection based on convolutional neural networks", Proc. Int. Conf. Syst. Sci. Eng. (ICSSE), pp. 101-104, Aug. 2021.