

# Abstract

The global Sports Analytics industry, valued at approximately USD 3 billion in 2023, is projected to experience a robust Compounded Annual Growth Rate (CAGR) of 20-30% over the next decade (Emergen Research, 2023; Markets and Markets, 2023). This paper summarizes various applications of big data in soccer, spotlights a successful case and projects a future outlook based on expert opinions.

Chapter 1 lays the groundwork talking about important concepts in soccer and big data to be understood to appreciate the following chapters.

Chapter 2 delves into the pervasive influence of big data in soccer, exploring the extent to which data technologies have shaped player performance, injury prevention, and strategic decision-making. The analysis underscores the transformative impact of big data analytics on the sport.

In Chapter 3, two compelling case studies including one centered on the Qatar World Cup 2022 highlight the strategic deployment of big data. The cases illuminate how sophisticated data analytics played a pivotal role in planning and organizing global sporting events, offering valuable insights into the intersection of big data and professional soccer.

Chapter 4 propels us into the future of big data in soccer. Discussions on player recruitment, injury management, and fan engagement reveal the dynamic landscape awaiting the sport. The analysis anticipates the continued evolution of big data applications, offering a glimpse into how technology will redefine the soccer experience for players, teams, and enthusiasts.

By exploring the current landscape, examining a real-world application in the Qatar World Cup, and projecting future developments, this paper provides valuable insights into the transformative power of big data in soccer.

# 1. Current Scenario | Big data in Soccer

## 1.1 Big data in sports

Big data has become increasingly significant in the field of sports, offering valuable insights, improving performance, enhancing fan engagement, and aiding in decision-making processes. Here are several ways in which big data is utilized in the realm of sports (Rein & Memmert, 2016):

**Performance Analysis:**

* Player Performance: Teams use data to analyze individual player performance, tracking various metrics such as speed, distance covered, heart rate, and more during training and actual games. This information helps coaches tailor training programs and make strategic decisions.
* Team Performance: Big data is used to analyze team dynamics, formations, and overall strategy. Teams can identify patterns and trends that contribute to success or failure in different game situations.

**Injury Prevention and Management:** Big data helps in monitoring players' physical condition and detecting signs of fatigue or potential injuries. This proactive approach allows teams to adjust training loads, manage recovery, and prevent injuries more effectively.

**Fan Engagement:** Sports organizations leverage big data to enhance fan engagement by providing personalized experiences. This includes tailoring content, promotions, and experiences based on fan preferences and behavior.

**Recruitment and Scouting:** Teams use data analytics to identify and recruit talent. Player statistics, performance data, and scouting reports are analyzed to make informed decisions about potential signings or draft picks.

**In-Game Decision Making:** Coaches use real-time data during games to make informed decisions. This includes insights into player performance, opponent strategies, and situational analysis, helping coaches make strategic substitutions or tactical adjustments.

**Wearable Technology:** Athletes often wear sensors and other wearable devices that collect data on their movements, biometrics, and overall health. This data helps in monitoring fitness levels, optimizing training routines, and ensuring peak performance.

**Sports Betting and Fantasy Sports:** Big data plays a significant role in sports betting and fantasy sports. Analyzing historical data, player statistics, and performance trends helps enthusiasts make informed decisions when drafting fantasy teams or placing bets.

**Venue Management:** Stadiums and sports facilities use big data for crowd management, security, and optimizing the overall fan experience. Data on attendance patterns, concessions, and traffic flow can be analyzed to improve operations.

**Broadcasting and Media:** Big data enhances the broadcasting and media aspects of sports by providing insights into viewer preferences. This information is used to tailor content, improve coverage, and optimize advertising strategies.

**Social Media Analysis:** Teams and athletes use big data to analyze social media trends and sentiments. This helps in shaping marketing strategies, understanding fan sentiment, and engaging with the audience more effectively.

The integration of big data analytics in sports has revolutionized the industry, offering a competitive edge and enriching both the sporting experience for fans and the performance of athletes and teams.

## 1.2 Real-time Football Analytics | Apache Spark

The landscape of football is currently undergoing a significant shift towards data-driven methodologies, with an unprecedented demand for real-time analytics. This endeavor delves into the realm of real-time football analytics leveraging Apache Spark, aiming to elevate match analysis, engage fans, and mitigate injuries. The research kicks off by establishing clear objectives and meticulously gathering data from diverse sources such as player tracking systems, social media platforms, and in-game statistics. At the core of this initiative is Apache Spark, a robust real-time data processing framework adept at handling extensive data volumes generated during live matches. This technology facilitates data transformation, feature extraction, and the application of machine learning algorithms to furnish real-time insights. Intuitive dashboards and interactive elements enrich the fan experience, providing live commentary and sentiment analysis during matches. The accuracy and efficacy of the real-time analytics undergo thorough scrutiny through comparisons with actual match outcomes and user feedback. By disseminating research findings and methodologies, this project makes a significant contribution to the progression of real-time football analytics. The potential impact is far-reaching, fundamentally transforming how football matches are analyzed, amplifying fan engagement, and furnishing invaluable insights for decision-makers in the football domain. This undertaking underscores the potency of Big Data technologies and a fervent commitment to the beautiful game in the dynamic landscape of football analytics (Talaviya et al., 2023).

## 1.3 Innovative analytics software for Soccer

Many software developers have sold their rights to soccer teams and management companies, thus not allowing the public to use them for sport analytics. However, there are some open-source software available which are used in soccer.

* **StreamTeam-Football:** STREAMTEAM-FOOTBALL allows to analyze football matches fully automatically and in real-time on the basis of tracked position data using a data stream analysis approach (Probst et al., 2020).
* **Footdata-pro:** Footdata-pro is being built with a set of open source technologies involved in building a multi-platform web based integrated football information system, supported in three main modules: user interfaces, databases, and the tactical plan detection and classification(Rodrigues et al., 2013).

## 1.4 Impact of Big data in Soccer

Soccer has evolved into a thriving industry, witnessing an unprecedented surge in financial influx as clubs strive to secure and maintain their top positions, transforming windfalls into consistent revenue streams. In this landscape, the integration of data and soccer analytics has become instrumental both on and off the field, mirroring its impact across various business sectors.

Within the UK, stadiums are adorned with cameras meticulously tracking players' movements, even when not in possession of the ball. Each game generates a staggering 1.4 million data sets, equating to 10 data points per second for every player. This wealth of information empowers managers, performance analysts, and backroom staff with insights that transcend the capabilities of the human eye, unraveling the intricate layers of on-field dynamics. Player monitoring extends beyond the pitch, encompassing GPS trackers, acceleration sensors, heart rate monitors, and comprehensive diet and sleep tracking, providing metrics that optimize training regimens and tactical strategies. This statistical approach not only aids in injury prevention but also forecasts recovery times and guards against overtraining. Beyond the pitch, every facet of a soccer player's life contributes to vast datasets.

While performance monitoring has been a staple in soccer, analytics play a pivotal role in reshaping the sport. Teams leverage data to scout and sign players strategically, enhancing their tactical advantage for upcoming seasons.

The emergence of data science in sports, popularized by Michael Lewis' Moneyball, is gaining prominence. Unlike sports with fixed positions and limited moves, applying predictive analytics to the fluid and dynamic game of soccer poses unique challenges. Data scientists collaborate with traditional scouts to identify the 'next big thing,' offering analytical approaches that uncover hidden talents in the transfer market.

Leading clubs, including Liverpool, have embraced data analysts like Michael Edwards, who revolutionized player recruitment methods. Analysts, working alongside traditional scouts, evaluate a player's entire performance history, enabling better judgments on long-term abilities and contributions to the club's success.

Performance data serves as a compelling justification for exorbitant expenditures in the transfer market. Real Madrid's acquisition of Gareth Bale for £85 million was supported by performance data, considering not only his on-field capabilities but also the forecasted profit of £41 million over six years through soccer shirt sales.

### 1.3.1 Spearheading analytics in sports using Football Manager

Prozone and similar companies spearhead performance analysis in sports, with most Premier League clubs utilizing their services. Integration with platforms like Football Manager has further enriched the player database, offering comprehensive information on various aspects such as dribbling, goal attempts, aggression, ball possession, and distance covered. This tool aids clubs in making well-informed decisions during transfer seasons, providing players from smaller clubs an opportunity to be noticed and signed by larger ones. While data analysis has significantly enhanced decision-making in soccer, there remains a need for the human element in assessing intangibles such as mentality and potential for growth. The evolving synergy between Big Data & Analytics and soccer holds immense promise, although there are still gaps to bridge on this exciting journey (Big Cloud, 2016).

# 2. Big data in Soccer | Challenges and Opportunities

Despite the revolutionary strides made using big data in soccer, there are several challenges that need to be addressed. Some of the key concerns are highlighted below (Bai & Bai, 2021):

**Anticipating Athlete Performance in Knowledge Graphs:**

Knowledge graphs are often used for predicting athlete performance based on past games. The present prediction models concentrate on feature extraction and machine learning algorithms. But the relationship between athlete performance and athletes, coaches, and events is often overlooked. These other factors are often very influential in soccer, given that coaches, field conditions, and teammates have great power over a player’s performance. To predict a player’s performance more accurately, the knowledge graphs should be derived from big data. Thus, constructing a knowledge graph that encompasses sports performance and related entities becomes a pivotal task. Additionally, the challenge lies in effectively utilizing the constructed knowledge graph to predict athletes' performance(Bai & Bai, 2021).

**Identifying Emerging Stars:**

The trajectory of a sports career hinges not only on an athlete's personal abilities but also on their team and country affiliations. Cultivating outstanding athletes demands substantial resources from teams or nations. "Rising stars" in sports refer to athletes in the early stages of their careers who show potential for future stardom, despite not currently outshining their peers. Discovering these rising stars offers constructive guidance for national fund investment and early support for athletes to display exceptional performances. The existing research predominantly employs statistical methods for athlete evaluations, leaving a gap in understanding how to construct a knowledge graph for identifying rising stars, presenting a challenging task (Bai & Bai, 2021).

**Shift in the revenue generation methods:**

“*Among the different aspects embedded in sports organizations’ management, “fan experience” has become one of the most relevant and innovative concepts developed. In the 1980s, ticketing represented 90% of the turnover; currently, ticketing accounts for between 15% and 20% of revenue. Revenue is now largely derived from other sources of entries: e.g. broadcasting, TV rights, sponsorship and commercial deals”*(Previati & Link to external site, 2020)*.* As big data shapes newer inventions in the soccer industry, many traditional revenue generating methods are becoming obsolete.

# 3. Case study on Big data and Soccer

## 3.1 Use of Big data analytics in Qatar World Cup 2022

At Qatar world cup 2022, players were sent individual performance analytics captured using advanced camera technology. (Adam, 2022) This may not seem like a big jump from the analytics that has been in place for a decade now. Training data analytics, player analysis for trade season etc. were all batch data processing. However, the use of streaming data from sources like motion sensers, precision GPS, wearable vests etc. being processed in real-time is a new development in soccer. With the analytics becoming more real-time, the role of analysts is also expected to become more active in the operations side – a position only for athletics-related people till date.

As opposed to some other games which have made these strides quite some time ago (like baseball), soccer is just now catching up to the level of technical sophistication through no fault of its own. While it may be thought to be the resistance of the old guard that prevented soccer from making these progresses before, the old guards reside in every game just like it did in baseball. The differentiating factor between games with different levels of analytics is mostly the nature of the sport, for example – soccer is an invasion game whereas baseball is a stop-start game.

Clubs buy information from commercial companies that code filmed matches to record outcomes of thousands of in-game events. In soccer, these events are dribbles, passes, shots, tackles etc. These data encoding efforts are increasingly moving towards computer-vision (a type of Artificial Intelligence).

Due to the fluid nature of the sport, soccer sport analytics is very expensive and does not yet produce reliable returns. Set-pieces (free-kicks, corner kicks, penalties etc.) are more stop-start in nature and owe themselves to relatively simpler analytics applications. Soccer clubs, even in the largest leagues, vary massively in their wealth. For the foreseeable future, it is expected that the richest clubs will invest heavily in closed-door proprietary systems and reap marginal benefits from on fluid in-game analytics.

## 3.2 Double Poisson Model-a revolutionary prediction method for Euro 2020

Inaugurated in 1982, the double Poisson model, which assumes that goals scored by each team follow a Poisson distribution with a mean contingent on attacking and defensive strengths, remains a popular method for forecasting football scores. Despite the emergence of newer methodologies, this study focuses on the pre-tournament predictions generated by this model for the Euro 2020 football tournament. Remarkably, these predictions emerged victorious in the Royal Statistical Society's prediction competition, underscoring the model's ability to yield high-quality results. The model introduces a range of novel analytical results precisely outlining the conditions for the existence and uniqueness of solutions to the equations governing the model parameters. Following these derivations, there were some potential issues with the model— the undue influence of results from weaker teams—demonstrating the efficacy of disregarding outcomes against the weakest opposition. It also contrasts the model's predictions with the actual results of Euro 2020, revealing its exceptional accuracy in forecasting the number of goals scored. The case delves into the choice of the start date for the dataset, demonstrating that the authors' decision to commence the dataset just after the previous major international tournament was nearly optimal, at least in this context. These findings contribute to a deeper understanding of the mathematical behavior of the double Poisson model, providing empirical support for its effectiveness as a match prediction tool. To assess the model's efficacy, the study employed it to generate pre-tournament predictions for the delayed Euro 2020 competition. These predictions were entered into the RSS Euro 2020 Predictor Competition for comparative evaluation with other models. Figures 8 and 9 present a comparison between the predicted and actual values of goals scored and conceded by each team, showcasing a close alignment between the predictions and the observed outcomes. Notably, the majority of the actual values fall within the 95% prediction intervals, emphasizing the model's precision, particularly in predicting goals scored (Penn & Donnelly, 2022) .

While the model's predictions closely match actual values, a notable observation is the potential overestimation of variance in goals scored, a common critique of the Poisson model. Despite this, the study highlights that this assumption does not significantly impact the accuracy of means, which consistently prove to be reliable predictions.

# 4. The future of big data in soccer

In the future, big data in soccer will revolutionize player performance, injury prevention, and fan engagement. Wearable tech and GPS systems will offer real-time insights into player movements, while predictive analytics will help predict and prevent injuries. Advanced scouting models and automation will transform player recruitment, and personalized fan experiences will be enhanced through augmented reality and virtual reality technologies. Financial decisions will be optimized, and VAR decisions refined through data analytics. Global collaboration in data sharing and ethical considerations in data use will shape the evolving landscape of soccer analytics. The future promises a dynamic and technologically advanced soccer experience for players and fans alike (*Geolocation Tracking in Sports The Future of Performance Analysis*, 2023).

# References

Adam, D. (2022). Science and the World Cup: How big data is transforming football. *Nature*, *611*(7936), 444–446. https://doi.org/10.1038/d41586-022-03698-1

Bai, Z., & Bai, X. (2021). Sports Big Data: Management, Analysis, Applications, and Challenges. *Complexity*, *2021*, e6676297. https://doi.org/10.1155/2021/6676297

*Geolocation Tracking in Sports The Future of Performance Analysis*. (2023). Utilities One. https://utilitiesone.com/geolocation-tracking-in-sports-the-future-of-performance-analysis

Penn, M. J., & Donnelly, C. A. (2022). Analysis of a double Poisson model for predicting football results in Euro 2020. *PLOS ONE*, *17*(5), e0268511. https://doi.org/10.1371/journal.pone.0268511

Previati, A. (2020). *The Importance of Fan Engagement and Fan Management in Sports* [M.B.A., St. John’s University (New York)]. https://www.proquest.com/docview/2467547004/abstract/BDE7FABB24EA49FBPQ/1

Probst, L., Schuldt, H., Seidenschwarz, P., & Rumo, M. (2020). StreamTeam-Football: Analyzing Football Matches in Real-Time on the Basis of Position Streams. *2020 IEEE International Conference on Big Data (Big Data)*, 637–644. https://doi.org/10.1109/BigData50022.2020.9377914

Rein, R., & Memmert, D. (2016). Big data and tactical analysis in elite soccer: Future challenges and opportunities for sports science. *SpringerPlus*, *5*(1), 1410. https://doi.org/10.1186/s40064-016-3108-2

Rodrigues, P., Belguinha, A., Gomes, C., Cardoso, P., Vilas, T., Mestre, R., & Rodrigues, J. M. F. (2013). Open Source Technologies Involved in Constructing a Web-Based Football Information System. In Á. Rocha, A. M. Correia, T. Wilson, & K. A. Stroetmann (Eds.), *Advances in Information Systems and Technologies* (pp. 715–723). Springer. https://doi.org/10.1007/978-3-642-36981-0\_66

Talaviya et al. (2023). BIG DATA ANALYSIS IN FOOTBALL. *International Research Journal of Modernization in Engineering Technology and Science*.