Name: Aaditya Pal

Rollno:**4045A029**

Class: TYBSCIT

Control id:2021080314

Topic: Big data in the sports sector

Abstract:

In today's world, data is generated at a massive scale and in every sector that man can think of. Sports is no different. Big data analytics involves the collection, storage, and analysis of large volumes of data to retrieve valuable insights and make decisions. This multi-billion-dollar industry uses big data analytics to predict match outcomes, enhance a team's and individual player's performance, and improve fan engagement. It presents large business corporations with the chance to use the sports market as a marketing platform and make better decisions.

Various technologies have emerged to make this possible such as Hadoop by Apache which provides a way to visualize the exabytes and zettabytes of data that is generated.

Introduction:

The era of big data has transformed the way things operate in major sports competitions such as Football World Cups, the Olympics, Cricket tournaments, and motorsports. New technologies such as Decision Review Systems(DRS), VAR (Video assistant referee), etc. have emerged for making better on-field decisions regarding fouls or determining whether a batsman is out. This process is also streamed live to the audience. To make this possible, huge amounts of data are transmitted and analyzed in real-time. This gives an idea of the possibilities when the capability of big data is leveraged. Big data has not only unlocked the ability to assess performance and scout for better talent but also has provided a treasure of historical data with videos.

To gather this data, the concept of data mining is used. Data mining refers to the extraction and identification of useful information from large sets of data. It is often termed as knowledge mining which relies stress on mining from vast sets of data. All these data are of no use until the useful information is extracted from it. Data mining requires data cleaning, data integration, data transformation, pattern evaluation, and data presentation [4]. The sheer amount of data generated has led to these organizations and teams having dedicated data analysis departments. Various mathematical procedures are used to analyze this data. The research work signifies the importance of Big Data in improving sports performance.

Big data analytics in sports can be divided mainly into three categories:

- The field-level analysis focused on the players, coaches, and teams
- Market and management analysis used by policymakers to make better decisions
- Analysis of the literature that uses sports data to answer various questions in the field of economics and psychology.

Big data extends outside the field as well. Big corporations analyze customer trends and ideologies to make better business decisions. For example, by analyzing social media data, teams can identify which players and events generate the most hype and use this information to create targeted marketing campaigns.

Detailed Review:

Review 1

Sports analytics and the big-data era

Detailed review

Elia Morgulev et al. [1] say that big data analytics in sports can be divided mainly into three categories:

- The field-level analysis focused on the players, coaches, and teams
- Market and management analysis used by policymakers to make better decisions
- Analysis of the literature that uses sports data to answer various questions in the field of economics and psychology.

- **Field-level oriented analysis**

  The first form of sports analytics was observed in the United States around the year 1960. It was carried out using Notational analysis. Notational analysis is a way of recording performance in a quantifiable manner. From the year 1990 onwards technological advancements and computerization introduced big data into sports.

  This was evident in basketball when the NBA distributed Advanced Scout Software to 16 NBA teams. The data collected was logged into the system. The information collected included the number of players' shot attempts, the type of shots taken, and the number of rebounds taken by players. This data was then uploaded to an electronic bulletin from where any of the participating teams could download the data. The Advanced Scout Software discovered various patterns from each game. Further down the line, In 2003-04, researchers discovered useful predictors of shot location and goal percentage. This led to the proposition of a new statistical model for analyzing basketball shot charts.

  The three-point shooting style evident today in the NBA is a result of the analysis that considered the shot location by which the analysts could develop a model of expected points from each location on the court.

  Similarly in English Premier League(A football league in the UK) made some of the data available to the fans for open-source analysis. A prime example of this can be the goalkeepers look at the information on probable directions of penalty shots, based on the shooters' previous statistics. Advanced Information Systems are used like Opta, Prozone, etc that provide the analyst with heat maps and visualizations of ball movements on the pitch.

  Also, GPS devices are another source of information for data on accelerations, changes of direction, movements, etc. Strength and conditioning coaches devise training regimes based on this objective data generated. The risk of injury can also be identified well in advance with the help of these GPS divides, accelerometers, gyroscopes, etc. [1]

- **Study of human behavior using sports data**

  One of the notable findings in psychological-economic research with the help of sports is the debunking of the theory that a player who is constantly putting the ball in the net produces more of the same as he attempts more shots. The analyzed data showed that there was no positive correlation between the outcomes of successive shots taken in basketball. The data was taken from 48 games of an NBA team and this was concluded from the research.

  Another study also observed the NBA games with a dataset of over 83,000 shots from the 2012-2013 season, combined with the data of both the players and the ball position in each shot attempt. The researchers were able to construct a comprehensive model of shot difficulty and this demonstrated that the players who exceeded their expectations in their recent shots were more confident in taking shots from a distance and facing a tighter defense.

Similarly, In football, penalty kicks were observed to examine von Neumann's Minimax Theorem(MSNE) which says that players may play some of their strategies with certain probabilities rather than pursue a single pure strategy, Penalty kicks in football provide a large set of data through which the theory was tested. [1]

- **Management and policymakers' decision-oriented analysis**

  Big data is used from a business perspective as well as it analyses the impact of the sports events such as deciding the optimal cost of a ticket. A detailed cost-benefit analysis by an organization could determine whether hosting the Olympic Games positively affected host cities' long-term growth, GDP, etc.

  Other examples include:

  - Baade and Matheson [1] analyzed the cost side of hosting the Olympic games concluding that they were a money-losing enterprise with the upside that the long-term benefits may include increased tourism, improvement of infrastructure, etc
  - The NFL implemented ticket revenue sharing, and equal broadcast revenue sharing based on the research that this creates a competitive balance between the teams leading to more unpredictable outcomes, which the audience finds quite entertaining
  - As for ticket pricing, the seating location was used to determine the cost. Later on, the price of the tickets started to fluctuate based on the changing market conditions. For eg: the Boston Red Sox baseball team monitored the behavior of the fans and optimized the location of memorabilia and the distance between the fans and the food stalls based on the entrance used by them.
  - Social media is analyzed to gather data about fan interaction. The analysis done for the top 5 EPL teams showed the human behavior patterns that influenced the social media marketing campaigns. [1]

Review 2

Big Data and Formula One

Detailed review:

Formula 1 is a sport where time plays a very crucial role. Even the smallest of differences can affect the result of the race massively.

There are a number of variables that need to be tweaked before any race weekend such as the tires should be at the optimal pressure and temperature in order to get the best of them, calculating the fuel consumption, tweaking the wings to get the required amount of downforce and what might be the best settings for that particular track. Even a minute change in these settings can have a cascading effect on other variables. Mercedes-AMG has been at the top of the sport for 8 consecutive seasons and they make use of big data analytics massively. They could collect around 500 GB of data [2] every race and around 10TB of data over the entire season. There are almost 200 sensors fitted to the car. [2]And this data is not only recorded by the teams but some of it is also available to the broadcasting services that use it to show real-time graphics. Data analytics has been at the forefront of making decisions to develop the car, devise the optimal strategy, to enhance the performance of the car at Mercedes-AMG. They have a dedicated data analytics team that analyzes every car design, every practice session, and every driver biometric in real time to gain insights as to what is the best decision to be made at that time. An example of this could be the 2021 Spanish Grand Prix where it was quite evident how the real-time telemetry data was used by Mercedes-AMG to win the race. The data included the rival's data, prediction of what the rivals RedBull do, radio messages, telemetry data, driver inputs, etc. all came together to result in a victory. And some of these communications were even broadcasted live. Data analytics is employed to achieve success even if it means saving a second here and there. [2]

Review 3

Leveraging Big Data Analytics Utilizing Hadoop Framework in Sports Science

Detailed review:

Hadoop, a Java framework has the capability of handling enormous, complex, and variety of data sets in a distributed environment. The system consists of thousands of nodes each running various applications and despite the failure of a few nodes, the system is available and functional as always. Hadoop takes inspiration from the Google MapReduce Algorithm.

How does MapReduce Work:

- The input data is first split and then a single map task is associated with each split that executes the concerned map function with that split.
- The time taken to process multiple splits is much less compared to that taken by processing the whole input and the load balancing is well-maintained due to parallel processing.
- Usually, the split size is the same as the HDFS(Hadoop file system) block size of 64MB.
- The map tasks store the results onto a local disk and don't write to HDFS as they are not the required output.
- If the node fails, then Hadoop makes use of another node and re-executes the map task.
- The results of the map task act as input to the reduce task.
- The output of the reduce task function is the required result and is hence stored in HDFS. The local node receives the first copy and other copies are placed in off-rack nodes. [3]

How Hadoop works with an example:

G. Jagdev et al. [3] have considered the following example. Suppose we need to find the lowest bowling average of the respective Indian bowlers from the year 1980 to 2017. Initially, this data is gathered into four different files. The goal is to find the most efficient bowler in this period. The different phases involved to find a solution would be:

- Map phase: The first step is to create key-value pairs of the bowler and his average like <k,v>. Eg: <Anil Kumbhle, 29.65>
- Combiner phase: Then the most efficient bowler is filtered out from each file using a simple searching algorithm.
  Eg: Min = the least bowling average. if(v(second bowler).average < Min) {Min = v(average);} else{proceed with checking;}
- Reducer Phase: Once the most efficient bowlers from each file are figured out, the same code is executed to find the most efficient bowler among them.

Implementation example:

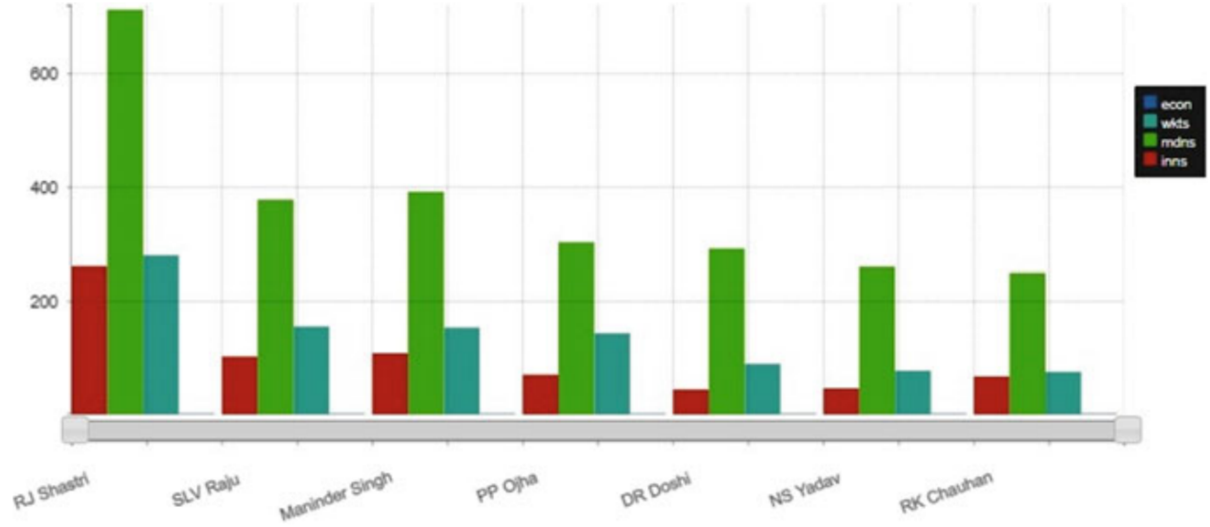
G. Jagdev et al. [3] carried out the following implementation to demonstrate the use of Hadoop. Consider the bowler data is stored in a structured database consisting of various attributes such as no.of matches, overs, maidens, runs conceded, average, etc. G. Jagdev et al. [3] made use of a Hadoop-based framework, Hortonworks Sandbox 2.2.0 in collaboration with VMware. They [3] made use of the Query Editor of Apache Hive to write queries for the same.

An example of a query:

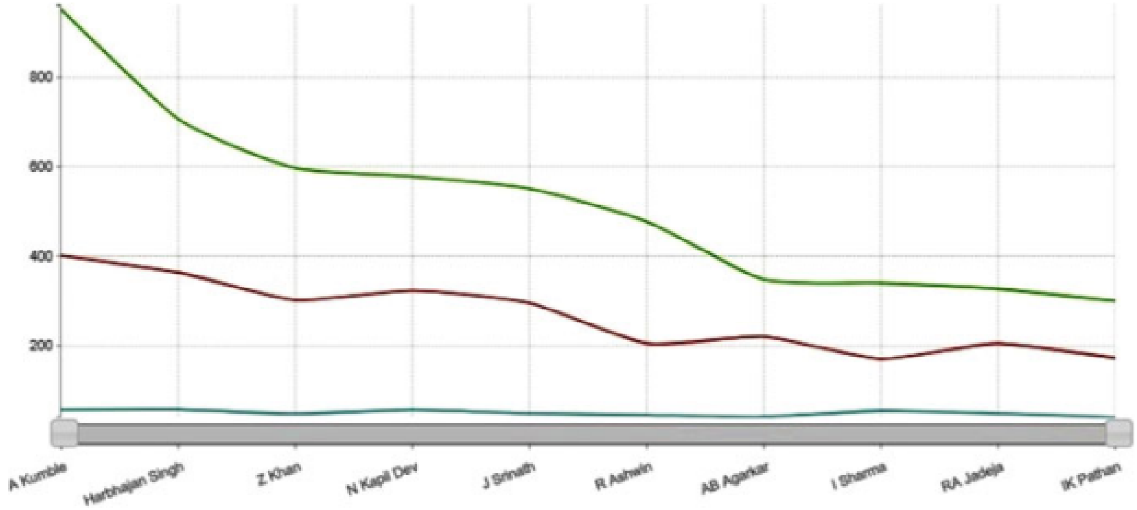
1. *Drop table if exists matches; create table matches as select player,mat,wkts,ave,ecom,sr from indian_bowlers where mat>=100 and wkts>=300;*

This gives a table consisting of bowlers who have played over 100 matches and taken 300 wickets in those matches.

2. *Drop table if exists bowlerperformance; create table bowlerperformance as select player,inns,overs,mdns,wkts,econ,mwi from indian_bowlers where overs>=1000 and econ<=3;*
This results in a table that consists of bowlers who have bowled over 1,000 overs and have an economy of less than 3.



The above figure displays the graphical outcome acquired from table *bowlerperfomance.* [3]



The above figure displays the graphical outcome acquired from table *matches.* [3]

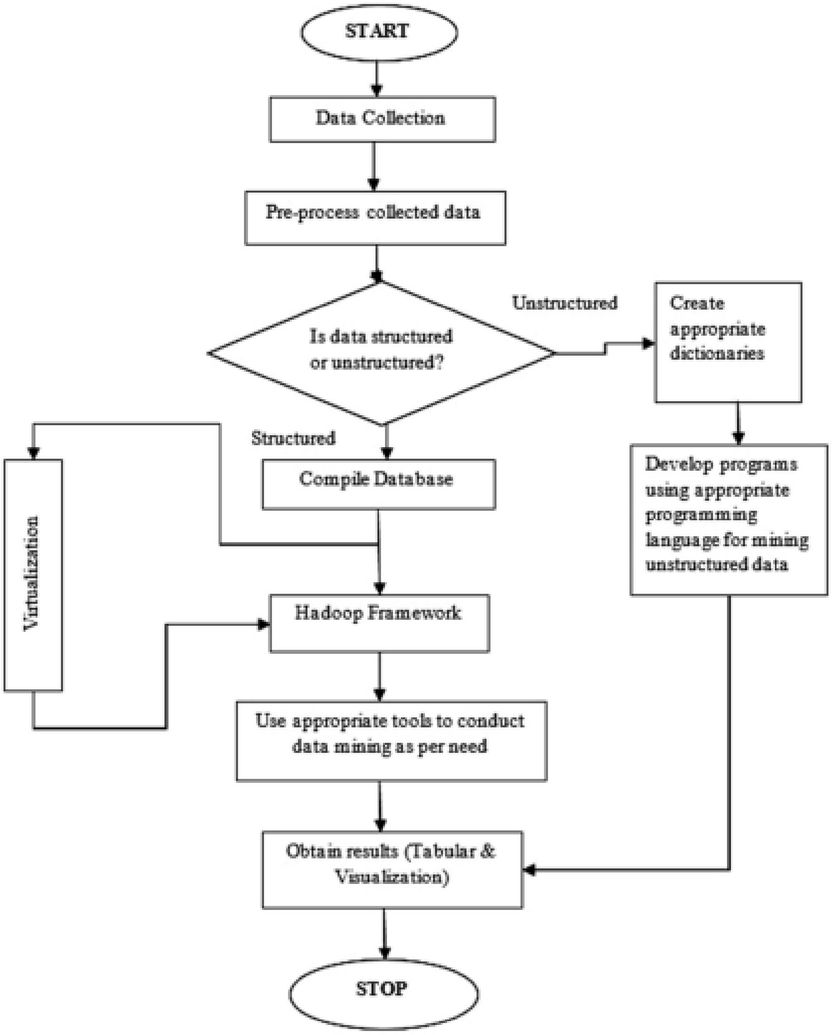
Review 4

Big data analytics in the sports sector

Detailed review:

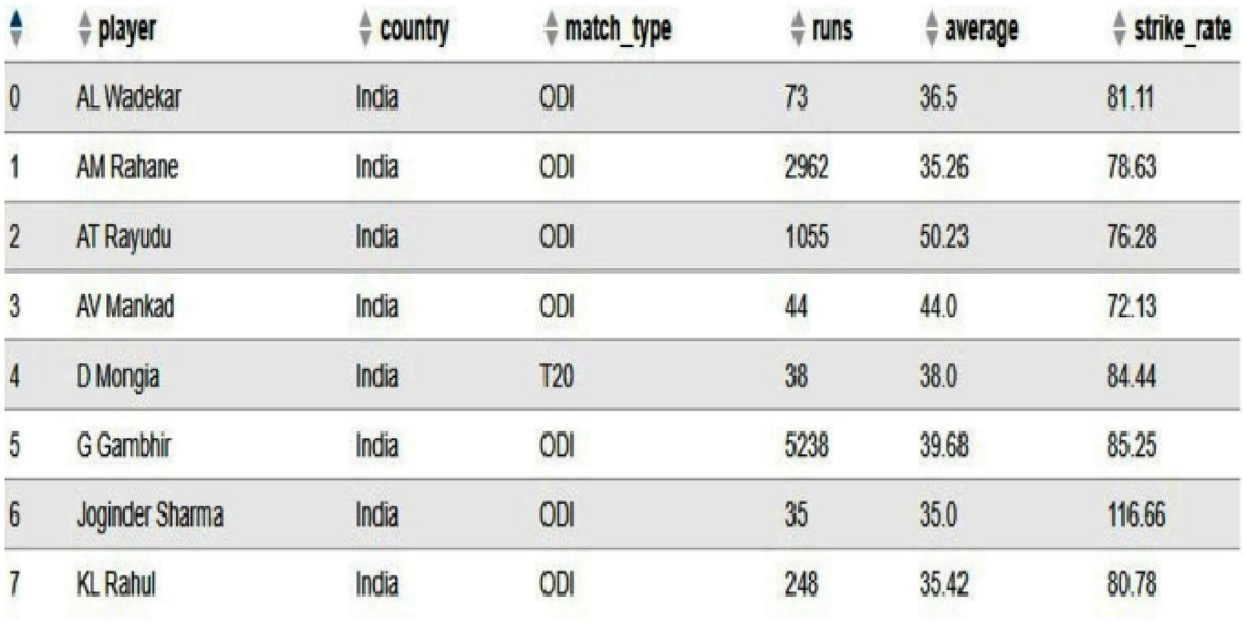
Amandeep et al. [4] say that after the data is mined and collected either from various websites like espncricinfo.com or other reliable websites, the data can be analyzed. The collection of data from reliable sources is very important because if there is an error in the source of the data the entirety of the analyzed data would be inaccurate and this defeats the main purpose of data analysis itself. Hence, after the data is collected the data can be stored in a relational database i.e. in a structured manner, or in an unstructured manner. The database of choice can highly influence the speed of the operations performed on it. Velocity, Volume, and Variety are the three main factors that are important in Big Data. In terms of sports, like cricket, for example, the details of the batsmen are retrieved and the data is represented graphically. An algorithm is used to mine the data effectively. This algorithm is vastly implemented in languages like Python and Java. Python receives vast support in terms of libraries available to represent data and analyze it and hence is used in most applications. An example of this can be as follows:

- The algorithm:


START
Data Collection
Pre-process collected data
Is data structured or unstructured?
Unstructured
Create appropriate dictionaries
Structured
Compile Database
Develop programs using appropriate programming language for mining unstructured data
Virtualization
Hadoop Framework
Use appropriate tools to conduct data mining as per need
Obtain results (Tabular & Visualization)
STOP

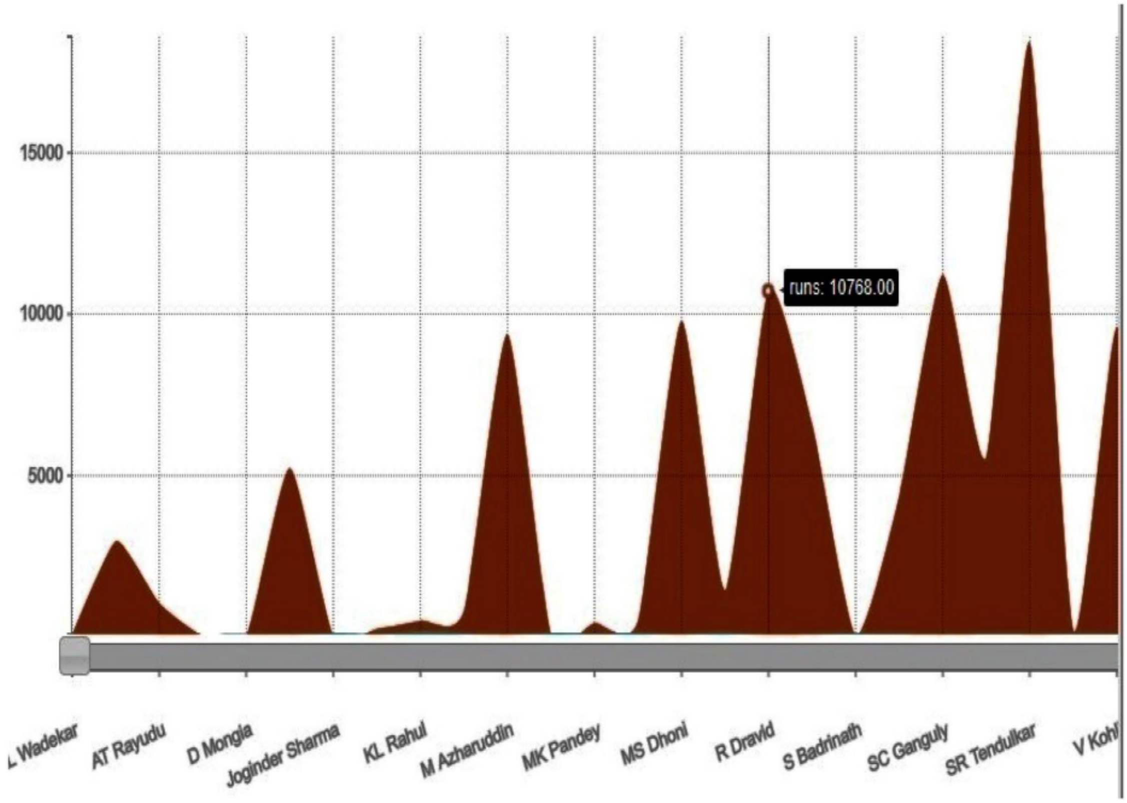
An overview of the algorithm [4]

- The database [4] :



	player	country	match_type	runs	average	strike_rate
0	AL Wadekar	India	ODI	73	36.5	81.11
1	AM Rahane	India	ODI	2962	35.26	78.63
2	AT Rayudu	India	ODI	1055	50.23	76.28
3	AV Mankad	India	ODI	44	44.0	72.13
4	D Mongia	India	T20	38	38.0	84.44
5	G Gambhir	India	ODI	5238	39.68	85.25
6	Joginder Sharma	India	ODI	35	35.0	116.66
7	KL Rahul	India	ODI	248	35.42	80.78

- They achieved the results [4]:



-

Review 5

Challenges of Big Data in Sports

Detailed review

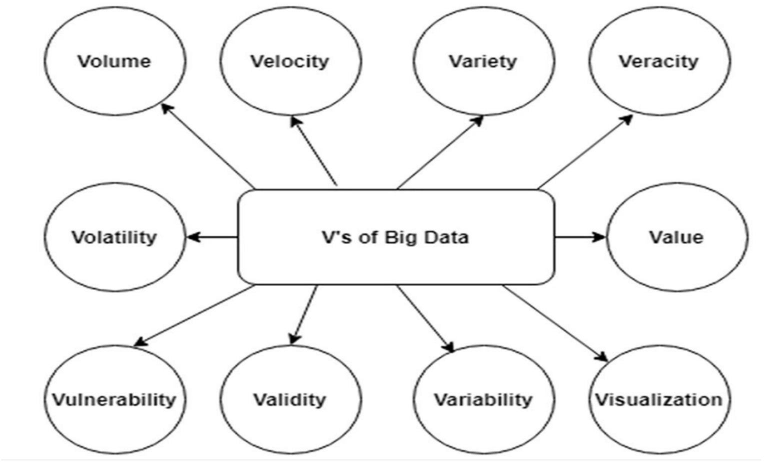


Figure 5.1 Challenges of Big Data [4]

Data Challenges

The challenges faced when dealing with big data can be divided into 10 different categories. These can also be called the V's of Big data. [4]

1. Volume:
   Big data itself has the word "big" referring to the vast amounts of data generated every second of every minute. EG: for a single IPL match alone, a few gigabytes of data are produced which includes the ball-to-ball analysis, video and audio analysis, decision review systems like ultra edge and hotspot player stats, fan interaction, revenue generated, weather forecast, etc.
2. Velocity:
   The speed with which the data is delivered becomes very crucial. For eg: In cricket, when a team opts for a decision to be reviewed, the data needs to be analyzed, and processed, and the graphics for the same need to be displayed. This involves the video, audio, etc. to be processed in order to determine whether the batsman is out. The fans involved in fantasy apps also depend on this data as they make their decisions based on it. Hence, the speed of data transmission becomes very important.
3. Variety:
   Data generated in sports is quite diverse. It includes the raw, unstructured [4], structured videos of the matches along with its audio, statistical data, fan interaction data, weather forecast data, revenue data, etc. Hence, processing this data becomes increasingly difficult
4. Veracity:
   It refers to the accuracy and precision of the data along with the risks involved with it. Every piece of information needs to be accurate as it drives important decisions that involve huge sums of money. Failure to do so can affect the trustworthiness, and the organization may lose a large number of fans as well. Similarly, a player's training regime may be incorrectly formed if the data received is incorrect, and instead of reducing the chances of injury, it may increase the risk of injury.
5. Value:
   Value refers to the worth of data being carried (*Amandeep Kaur et al.*) [4]. Worth is determined based on its effectiveness in multiplying and gaining profits.
6. Visualization:
   The data being generated also needs to be presented in a manner that it can be understood and here is where data visualization comes into play. The target audience should be able to use

the data generated in order to increase its effectiveness. New ways of representing data like clustering, sunbursts, tree maps, or network diagrams are required. [4]

7. Variability:

   Variability refers to a number of things. Data inconsistency, unpredictable speed, etc. come under variability. The discrepancies in the data need to be dealt with because the data is generated from various sources and in various formats.
8. Validity:

   It is quite similar to data veracity referring to the accuracy of the data. Ensuring validity requires effective data governance practices that directly affect the quality of the data, which is the need of the hour.

   Eg: A group of analysts found evidence of match rigging in professional sumo after finding discrepancies in the data between the years January 1989 and 2000
9. Vulnerability:

   It refers to the security concerns related when dealing with big data. Big data involves huge volumes of data and any security breaches are major ones due to this.
10. Volatility:

    It refers to how the data is stored. Big data needs to be retrieved and stored quickly and the technology needs to deal with this requirement of speed and volume. Performance is a major concern here.

Process Challenges

According to *Naseer T et al.* [5], the process challenges are the challenges that occur while processing big data. They are as follows:-

1. Data acquisition and recording:- All the data that is generated is not useful and the challenge is to separate what's important and what to discard. Such algorithms are required that process and reduce the data before storing it. Also, systems that can automatically generate metadata are required as currently there are no such systems and metadata is important as it describes the data and how it was recorded or measured.
2. Information extraction and cleaning:- The collected data comes in different varieties and can't be processed directly by just collecting it. The challenge is to extract the required data and structure it in a format that then can be processed. Moreover, data might not be always accurate hence it needs to be verified to ensure its quality
3. Query processing, data modeling, and analysis:- The challenge is the speed at which the data is exported out of the database in order to analyze it and then store it again to meet the real-time nature of the demands. The queries slow down as the data sets get larger.
4. Data integration and aggregation: Database design plays an important role during data analytics and hence needs to be designed by experts or someone with an abundance of knowledge in the particular domain for which the database is being developed.
5. Interpretation:- The challenge is to present the complex data in a manner that can be understood by the decision-makers. Hence systems with rich visualizations are crucial. [5]

Management challenges

According to *Naseer T et al.*[5], these challenges deal with legal issues regarding access to data. They are as follows:-

1. Security: - Securing large volumes of data is a huge challenge and the attackers can gain access to business secrets, customer data, etc. if the systems storing the data are not secured properly.
2. Privacy:- An individual's data can be extracted from various resources due to the generation of data from various sources which raises concerns regarding the privacy of the user.

3. Governance:- It refers to the data quality issues. If the data is of high quality, then the decision-makers can trust and rely on it to make better decisions. [5]

Conclusion:

In conclusion, Big data plays a very important role in today's world of sports. It determines which players to include in the team depending on the conditions, which players to scout for, which player lacks in which particular department, how to improve the team's performance, etc. Not only this but fantasy game users also use data to make better decisions. Big corporations also monitor fan engagement in order to develop better business strategies.

By looking at analysis carried out in sports such as basketball and football, it is understood that data analytics has been around for some time and has evolved through time. The data generated was used to modify player training regimes and proactively prevent injuries.

Motorsports is not immune to the explosion of big data. Formula One, the pinnacle of motorsports massively integrates big data and its analytics to achieve success in a sport where the winner is separated from the rest by just a matter of a few milliseconds.

Psychological and economic studies also gain outcomes from studying the data to understand human behavior.

In order to visualize and analyze the generated data, various technologies are present in the market. Hadoop is one of them that is widely used. Hadoop by Apache, is capable of handling complex data sets which constitute a variety of data in a distributed environment. It takes inspiration from Google's MapReduce Algorithm to find meaningful insights from the data.

Lastly, it's not all sunshine and rainbows for big data as it brings its own set of challenges. Dealing with massive amounts of data, which are generated in numerous formats at a very high rate is not easy. Newer technologies are in great demand that can fulfill these requirements.

Overall, Big data has opened up another dimension of possibilities in sports and this will very quickly become the norm and will continue to push the boundaries of what is possible.

Future Scope:

As wearable technology continues to grow at a rapid rate, the data that is collected will become more accurate. This would result in the rules being more strictly followed. For example, in the future, cricket bats and other equipment might be fitted with sensors that detect if the ball makes any contact with the equipment. This would result in accurately determining whether the ball hit the particular equipment or not. The wearables also would help to track the body vitals. This data can then be used to study how a human body copes with the pressure, what kind of pressure the player goes to, and how to perform better under such situations. The big data era has just begun and it will open countless other possibilities.

References:

[1] Sports analytics and the big-data era by Elia Morgulev, Ofer H. Azar, Ronnie Lidor
https://doi.org/10.1007/s41060-017-0093-7

[2] The Intertwine of Brain and Body: A Quantitative Analysis on How Big Data Influences the System of Sports

https://doi.org/10.1007/s40745-019-00239-y

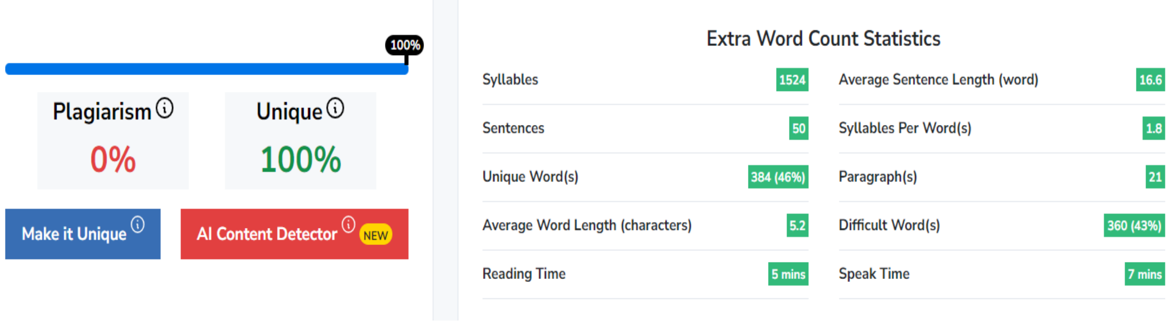
[3] Leveraging Big Data Analytics Utilizing Hadoop Framework in Sports Science
https://doi.org/10.1007/978-981-13-6295-8_22

[4] Analyzing and exploring the impact of big data analytics in the sports sector
https://doi.org/10.1007/s42979-021-00575-y

[5] Big Data Challenges by Nasser T* and Tariq RS
http://dx.doi.org/10.4172/2324-9307.1000135
https://www.researchgate.net/profile/Tariq-Soomro-2/publication/282281171_Big_Data_Challenges/links/560a53c908ae4d86bb137402/Big-Data-Challenges.pdf

Plagiarism reports

1. **Review 1**



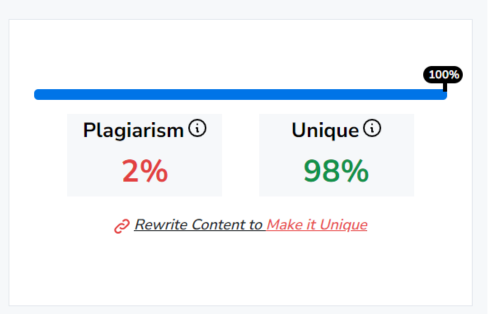
Plagiarism ⓘ 0%

Unique ⓘ 100%

Extra Word Count Statistics

Syllables	1524	Average Sentence Length (word)	16.6
Sentences	50	Syllables Per Word(s)	1.8
Unique Word(s)	384 (46%)	Paragraph(s)	21
Average Word Length (characters)	5.2	Difficult Word(s)	360 (43%)
Reading Time	5 mins	Speak Time	7 mins

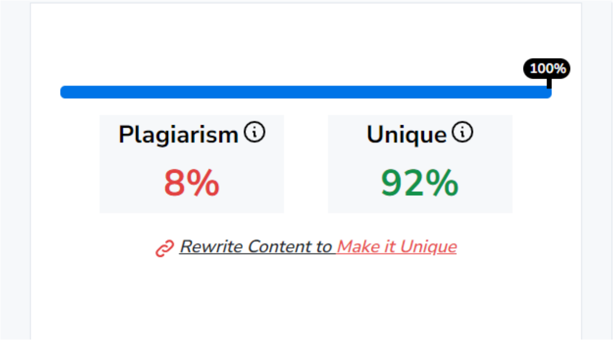
2. **Review 2**



Plagiarism ⓘ 2%

Unique ⓘ 98%

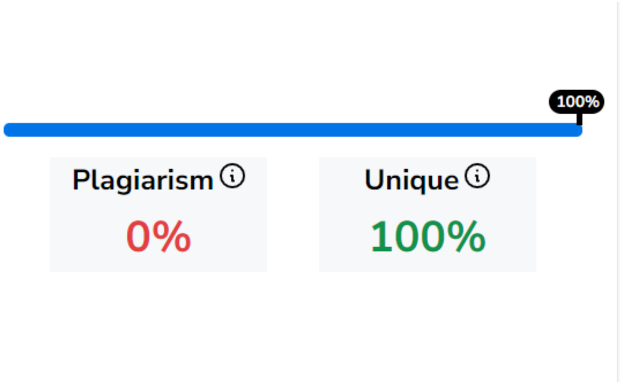
3. **Review 3**



Plagiarism ⓘ 8%

Unique ⓘ 92%

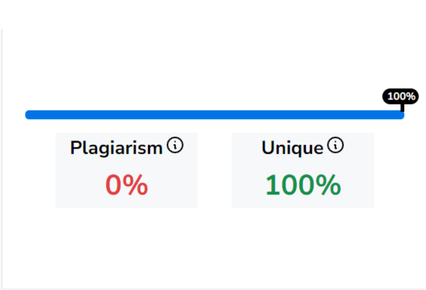
4. **Review 4**


100%

Plagiarism ⓘ 0%

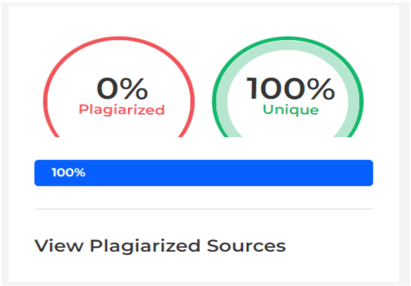
Unique ⓘ 100%

5. **Review 5**


100%

Plagiarism ⓘ 0%

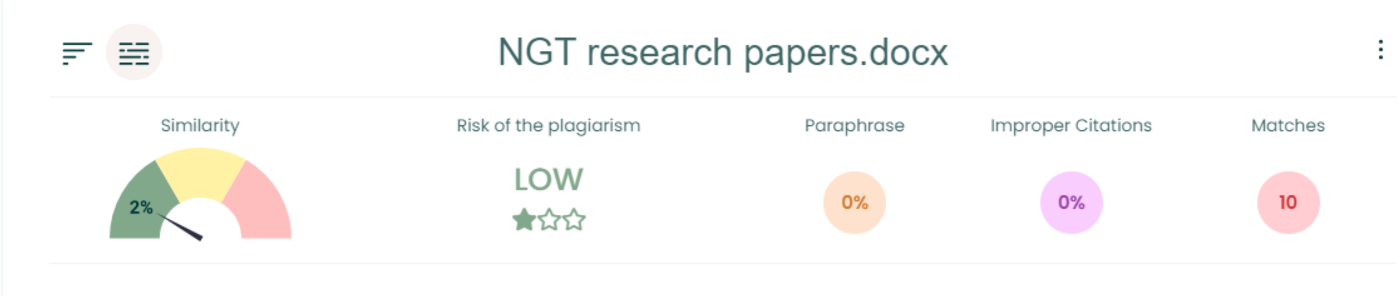
Unique ⓘ 100%


0% Plagiarized

100% Unique

100%

View Plagiarized Sources


NGT research papers.docx

Similarity	Risk of the plagiarism	Paraphrase	Improper Citations	Matches
2%	LOW ★☆☆	0%	0%	10