International Journal of Research Studies in Computer Science and Engineering (IJRSCSE)
Volume 4, Issue 4, 2017, PP 36-42
ISSN 2349-4840 (Print) & ISSN 2349-4859 (Online)
DOI: http://dx.doi.org/10.20431/2349-4859.0404005
www.arcjournals.org

Scrutinizing and Executing the Positive Aspects of Big Data in World of Sports via Apache Hadoop Framework

Sarabjeet Kaur[1], Dr. Gagandeep Jagdev[2], Dr. Ashwani Kumar[3]

[1]*Research Scholar (Ph.D), Punjabi University, Patiala (PB)*
[2]*Dept. of Comp. Science, Punjabi University Guru Kashi College, Damdama Sahib (PB)*
[3]*Associate Professor (ECE), Yadavindra College of Engineering, Talwandi Sabo (PB)*

Corresponding Author:** ***Dr. Gagandeep Jagdev, *Dept. of Comp. Science, Punjabi University Guru Kashi College, Damdama Sahib (PB). India*

Abstract: *Big Data is a prevalent term among data analyst involved in mining huge databases to extract hidden information. Any data which poses a challenge for currently existing database technologies is termed as Big Data. Today only big giants in the different fields like sports, retail, medical, stock exchange etc., can think of handling it because of expensive and huge infrastructure involved in handling it. This research paper is based on highlighting how sports science is benefited with Big Data. Each day many international and national matches are played in different sports and with each match new records, statistics, equations, and data is generated. With the arrival of wearable devices, the data collection has become very convenient for the analysts. This data assists the team management and coach to take critical decisions like which player should be in playing team and which should be rested, deciding the positions of the players, deciding the task to be fulfilled by the player, and much more. A team game can be improved by monitoring that whether players are playing the game as a team or does there any gap persists in the flow of the play. For handling such a huge database of thousands of players, analysts make use of powerful framework like Apache Hadoop. Apache Hadoop is one such framework which is capable of handling huge databases via its several components. It makes use of Map Reduce algorithm. The central theme of the research paper is elaborating the working of Map Reduce technology and Apache Hadoop framework in handling the enormous databases relevant to sports.*

Keywords: *Apache Hadoop framework, Big Data, MapReduce, sports.*

Abbreviations: *CSV, HDFS*

1. INTRODUCTION

"Big data"– which admittedly means many things to many people – is no longer confined to the realm of technology. Big data has thoughtful impact on the daily lives of consumers. Big data have successfully altered the approach businesses operate. The sports world is no longer an exception and have been benefited with the use of big data. Coaches excellently watch video to measure opponents' skills and advance their own players' performance, using knowledge gathered in the training room, but now they're also using big data analytics to gain an edge. With the use of sensors, GPS trackers, and RFID, training and coaching staff can capture information, feed the data into analytical engines and use it to inspire strategic decision-making. This approach can help them choose exactly the right player for any given play**:** As wearables like accelerometers, heart rate monitors and other devices have become more refined and precise, coaches and trainers are using them to keep players physically fit as they train and to restrain injuries, both during training and in the heat of competition. Wearables can also permit trainers and coaches to gather more accurate data on speed, acceleration and durability during combines and tests. This can help them accumulate more operative teams and choose the most promising players at draft time. The ability to produce and study data in real time to notify replay commentary, offer game-day facts and provide up-to-date player statistics can permit broadcasters to do a healthier job of interacting with fans and keeping them well-versed. Eventually, the goal for trainers, athletes, broadcasters, coaches, and others involved in sports decision-making is to influence real-time data to advance live performance. Like smaller companies that must find an edge to compete successfully with larger enterprises, a team with a smaller budget or more limited pool of athletes can use big data and analytics to gain an advantage. In that way, big data can make sports smarter.

2. MAJOR CONCERNS IN BIG DATA

The major challenges related to big data are mentioned as under and depicted in Fig. 1.

Data Volume – It refers to the massive quantity of data that is being produced every second, every minute and every hour of the day. 571 websites are shaped in a single minute. A total of 625000 GB of data is transmitted from one end to another in a single internet minute, may be terms of mails, pictures, posts etc. If we burn the amount of data present on planet earth today on DVDs and pile them in the form of a stack one upon another, the pile will be such huge that one can climb it and touch the moon, come back to earth and again repeat this process once [10, 11].

Data Velocity – Data are being generated at such high velocity that companies are finding it difficult to cope up with such high speed. They have to establish their infrastructure in such a manner that it is capable of handling such generated data. Social media and E-Commerce has rapidly increased the speed and richness of data used for different business transactions [8].

Data Variety - All the data being generated is totally diverse, consisting of raw, structured, semi structured and even unstructured data which is difficult to be handled by the existing traditional analytic systems. Mismatched data formats and data structures represent significant challenges that can lead to analytic collapse [10, 11].

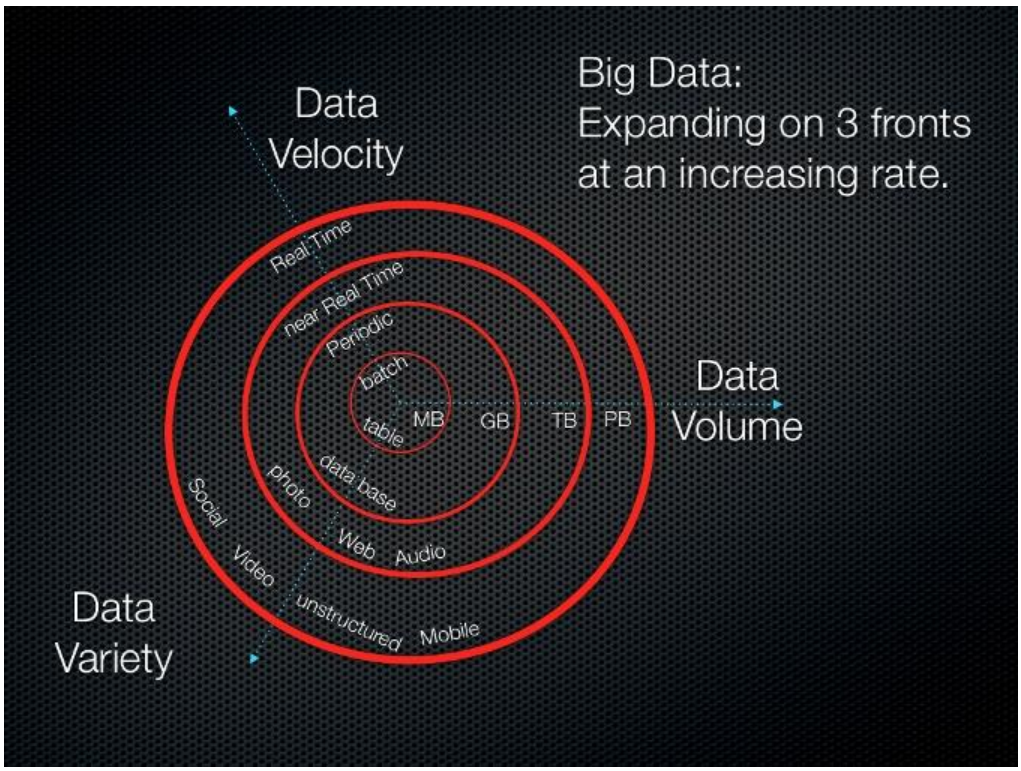


Fig1. *The figure depicts the 3 major challenges relevant to big data*

In addition to these, the three other concerns related o big data are mentioned below.

Data Value –There is a huge gap in between the business leaders and the IT professionals. The main concern of business leaders has been to just add value to their business and to maximize their profit. On the other hand, IT leaders deal with technicalities of the storage and processing.

Data Complexity – The biggest complexity faced while running big data using relational databases is that they require parallel software running on hundreds of servers and data scientists have to match and transform data across systems coming from various sources [10, 11].

Data Veracity - Veracity refers to the preciseness of data or how much faith, one can have on data. The data on the internet is not always accurate or precise. For example, if some male pretends himself as a female on his Face book profile, there is no authenticity check in such cases. Similarly twitter makes use of abbreviations and hash tags, but the big data enables us to work with even this type of imprecise data [1, 6, 9].

3. WORKING OF APACHE HADOOP FRAMEWORK AND MAPREDUCE

Hadoop is an Apache open source framework which allows distributed processing of large datasets across groups of computers written in Java using simple programming models. Hadoop is designed to gauge up from single server to thousands of machines, each having the capability of local computation and storage.

Hadoop framework comprises of different components mentioned as under.

- Hadoop Common – These comprise Java utilities and libraries vital for other Hadoop modules. These libraries offer file system and operating system level abstractions and encompasses the required Java files and scripts to start Hadoop.
- Hadoop YARN – This component is responsible for performing cluster resource management and job scheduling.
- Hadoop Distributed File System (HDFS) – This is a distributed file system accountable for providing high-throughput access to application data.
- Hadoop Map Reduce – This is the technology responsible for conducting parallel processing of large data sets.

To understand the Hadoop frame work, let's suppose entire system as a Hadoop zoo as shown in Fig. 2.

Fig2. *Hadoop Zoo*

- HDFS – HDFS are distributed cages where all animals live i.e. where data reside in a distributed format.
- Apache HBase – It is a smart and large database.
- Zookeeper- Zookeeper is the person responsible for managing animals play.
- Pig – Pig allows to play with data from HDFS cages.
- Hive- Hive allows data analysts play with HDFS and makes use of SQL.
- HCatalog helps to upload the database file and automatically create tables for the user.

4. IMPLEMENTATION AND CONTRIBUTION

The research work is primarily concerned with Indian cricket. A primary database has been constructed encompassing 14 different attributes listed below.

- *Player* – Refers to Cricketer name
- *Match_Type* – Refers to Type of match played(ODI/T20)
- *Span* – Refers to Cricketer career time period
- *Matches* – Refers to matches played by cricketer
- *Inns* – Refers to Number of innings played by the cricketer
- *Not_Outs* – Refers to number of times player remained unbeaten
- *Runs* – Refers to runs scored by cricketer in any particular format
- *Highest_Score* – Refers to the highest score of the cricketer in a particular format
- *Average* – Refers to batting average of cricketer in a particular format
- *Balls_Faced* – Refers to total balls palyed by cricketer in a particular format
- *Strike_Rate* – Refers to Strike Rate of the cricketer in any particular format
- *100* – Refers to centuries scored by cricketer in a particular format
- *50* – Refers to the number of half-centuries scored by cricketer in particular format
- *Ducks* – Refers to the cricketer's number of dismissals on zero.

The sight of the created database has been portrayed in the Fig. 3.

	Player	Match_Type	Span	Matches	Inns	Not_Outs	Runs	Highest_Score	Average	Balls_Faced	Strike_Rate	100	50	Ducks
2	SR Tendulkar	ODI	1989-2012	463	452	41	18426	200*	44.83	21367	86.23	49	96	20
3	SC Ganguly	ODI	1992-2007	308	297	23	11221	183	40.95	15235	73.65	22	71	16
4	R Dravid	ODI	1996-2011	340	314	39	10768	153	39.15	15126	71.18	12	82	13
5	M Azharuddin	ODI	1985-2000	334	308	54	9378	153*	36.92	12669	74.02	7	58	9
6	MS Dhoni	ODI	2004-2017	283	246	66	9101	183*	50.56	10284	88.49	9	61	8
7	Yuvraj Singh	ODI	2000-2017	293	268	38	8447	150	36.72	9665	87.39	14	51	18
8	V Sehwag	ODI	1999-2013	241	235	9	7995	219	35.37	7655	104.44	15	37	14
9	V Kohli	ODI	2008-2017	179	171	25	7755	183	53.11	8544	90.76	27	39	10
10	SK Raina	ODI	2005-2015	223	192	35	5568	116*	35.46	5938	93.76	5	36	14
11	A Jadeja	ODI	1992-2000	196	179	36	5359	119	37.47	7677	69.8	6	30	10
12	G Gambhir	ODI	2003-2013	147	143	11	5238	150*	39.68	6144	85.25	11	34	11
13	RG Sharma	ODI	2007-2016	153	147	23	5131	264	41.37	6077	84.43	10	29	10
14	NS Sidhu	ODI	1987-1998	136	127	8	4413	134*	37.08	6329	69.72	6	33	7
15	K Srikkanth	ODI	1981-1992	146	145	4	4091	123	29.01	5702	71.74	4	27	11
16	N Kapil Dev	ODI	1978-1994	225	198	39	3783	175*	23.79	3979	95.07	1	14	13
17	DB Vengsarkar	ODI	1976-1991	129	120	19	3508	105	34.73	5179	67.73	1	23	7
18	RJ Shastri	ODI	1981-1992	150	128	21	3108	109	29.04	5089	61.07	4	18	5
19	SM Gavaskar	ODI	1974-1987	108	102	14	3092	103*	35.13	4966	62.26	1	27	8
20	S Dhawan	ODI	2010-2017	76	75	3	3090	137	42.91	3435	89.95	9	17	2

Fig3. *The figure shows the glimpse of the constructed database under study*

The major objective of the research work is to perform mining of the created database in order to obtain desired results. The scripts can be written and queries can be passed to obtain information like

- Which cricketer has the maximum average in any particular format of the game
- Which cricketer has highest individual score against his name in any format
- Which cricketer has scored maximum centuries in any format of the game
- Which cricketer remained not out the maximum number of times
- Which cricketer enjoys the honor of playing the maximum number of matches in any format and much more.

The results can be obtained in the form of CSV (comma separated value) files and also in a graphical format like bar graphs, histograms, lines, and pie-charts. An example is mentioned as under to illustrate this concept.

Script1

Drop table if exists sports1;

Create table sports1 as select

Player,

Match_Type,

Span,

Matches,

Inns,

Not_Outs,

Runs,

Highest_Score,

Average,

Balls_Faced,

Strike_Rate,

100,

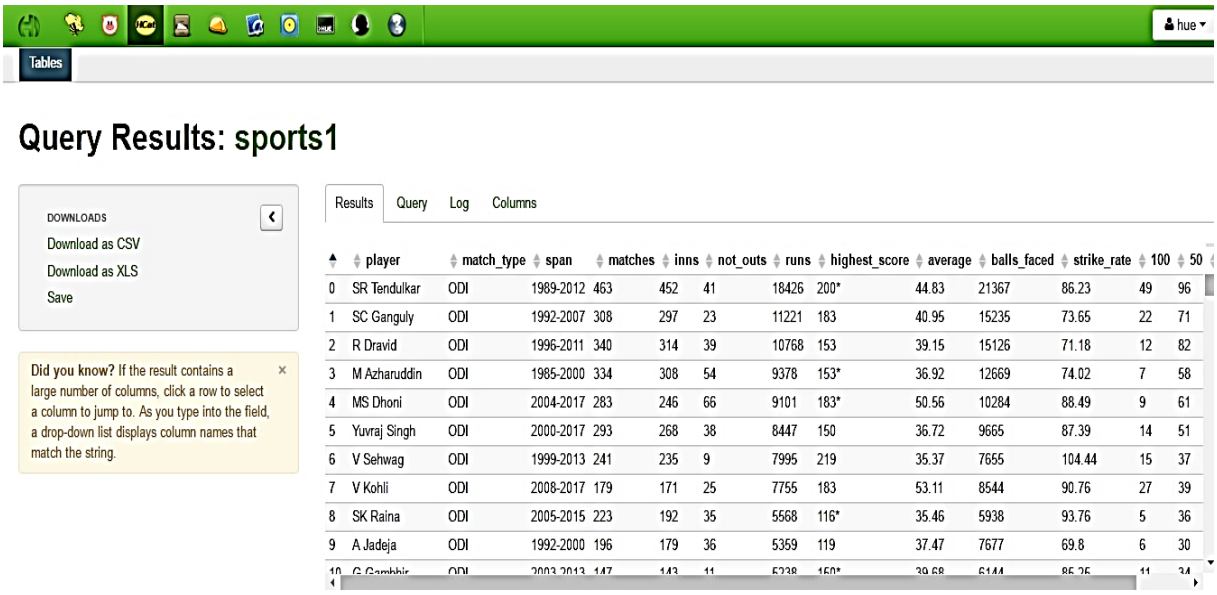
50,

Ducks

From cricketdb where Match_Type="ODI" and Matches>=50;

In the above mentioned query, *"sports1"* refers to the newly constructed table and the *"cricketdb"* refers to the name of the table already derived from the database.

After running above mentioned script, a table titled *"sports1"* is created which is shown in Fig. 4.



	player	match_type	span	matches	inns	not_outs	runs	highest_score	average	balls_faced	strike_rate	100	50
0	SR Tendulkar	ODI	1989-2012	463	452	41	18426	200*	44.83	21367	86.23	49	96
1	SC Ganguly	ODI	1992-2007	308	297	23	11221	183	40.95	15235	73.65	22	71
2	R Dravid	ODI	1996-2011	340	314	39	10768	153	39.15	15126	71.18	12	82
3	M Azharuddin	ODI	1985-2000	334	308	54	9378	153*	36.92	12669	74.02	7	58
4	MS Dhoni	ODI	2004-2017	283	246	66	9101	183*	50.56	10284	88.49	9	61
5	Yuvraj Singh	ODI	2000-2017	293	268	38	8447	150	36.72	9665	87.39	14	51
6	V Sehwag	ODI	1999-2013	241	235	9	7995	219	35.37	7655	104.44	15	37
7	V Kohli	ODI	2008-2017	179	171	25	7755	183	53.11	8544	90.76	27	39
8	SK Raina	ODI	2005-2015	223	192	35	5568	116*	35.46	5938	93.76	5	36
9	A Jadeja	ODI	1992-2000	196	179	36	5359	119	37.47	7677	69.8	6	30

Fig4. *Figure shows the created table titled "sports1" after running the script*

After the table *"sports1"* is created, the relations among the table attributes can be analyzed. Fig. 5 shows the snapshot showing the *"player"* against x-axis and *"strike_rate", "100",* and *"50"* attributes against y-axis.

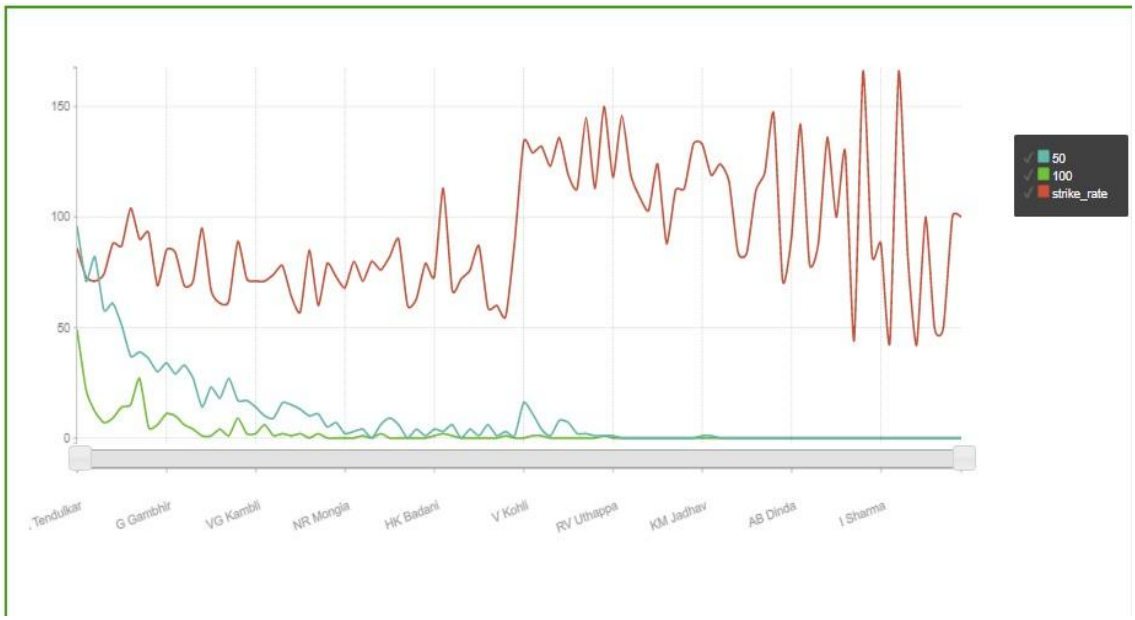


Fig5. *Figure depicts the snapshot showing co-relation between "player" and his "strike_rate", "100s" and "50s" scored*

Similarly, by writing appropriate scripts and queries, the desired results can be obtained from the database in tabular as well as in graphical form.

5. CONCLUSION

There is no doubt in the fact that Big Data has been a dramatic game changer in the world of sports, but it should be for informative purpose only. If a team halts believing on its expertise and gut instincts, the enthusiasm attached with the game would vanish. However, the use of Big Data can assist in taking critical decisions, constructing innovative policies and plans, forming the best team out of available players, and enhance the level of competition among the teams irrespective of the sports.

REFERENCES

[1] Dean, Jeffery, and Ghemawat Sanjay. 2004. "MapReduce: Simplified Data Processing on Large Clusters." Google.

[2] Katal, A., Wazid, M., &Goudar, R. H. (2013). Big Data: Issues, Challenges, Tools and Good Practices. IEEE, 404-409.

[3] Gagandeep Jagdev et al., "Analyzing and Scripting Indian Election Strategies using Big Data via Apache Hadoop Framework", *IEEE Xplore*, DOI: 10.1109/WECON.2016.7993431, INSPEC Accession Number: 17061464, 27 July 2017.

[4] Gagandeep Jagdevet al.," Scrutinizing Elections Strategies by Political Parties via Mining Big Data for Ensuring Big Win in Indian Subcontinent", *4th Edition of International Conference on Wireless Networks and Embedded Systems.*

[5] Gagandeep Jagdev et al., "Big Data proposes an innovative concept for contesting elections in Indian subcontinent", *International Journal of Scientific and Technical Advancements (IJSTA),*Volume 1, Issue 3, pp. 23-28, 2015, ISSN No. 2454-1532.

[6] Gagandeep Jagdev et al., "Comparing Conventional Data Mining Algorithms with Hadoop based Map-Reduce Algorithm considering elections perspective*", International Journal of Innovative Research in Science and Engineering (IJIRSE),* ISSN: 2454-9665 (O), ISSN: 2455-0663(P), Volume – 3, Issue – 3, March 2017.

[7] Gagandeep Jagdevet al., "Association of Big Data with Map-Reduce Technology Augments for Economic Growth in Retail", *International Journal of Engineering Technology Science and Research (IJETSR),* ISSN: 2394 - 3386, Volume 4, Issue 2, February 2017.

[8] N. Marz and J. Warren. Big Data: Principles and best practices of scalable realtime data systems. Manning Publications, 2013.

[9] R. Smolan and J. Erwitt. The Human Face of Big Data.Sterling Publishing Company Incorporated, 2012.

[10] Gagandeep Jagdevet al., "Analyzing Maneuver of Hadoop Framework and MapR Algorithm Proficient in supervising Big Data", *International Journal of Advanced Technology in Engineering and Science (IJATES)*, ISSN – 2348-7550,Volume – 05, Issue – 05, May 2017.

[11] Gagandeep Jagdev et al., "Implementation and Applications of Big Data in Health Care Industry", *International Journal of Scientific and Technical Advancements (IJSTA)*, ISSN: 2454-1532, Volume 1, Issue 3, pp: 29 – 34.

AUTHOR'S BIOGRAPHY

Dr. Gagandeep Jagdev, is working in the capacity of a faculty member in Dept. of Computer Science, Punjabi University Campus Guru Kashi College, Damdama Sahib (PB). His total teaching experience is above 11 years and has above 114 international and national publications in reputed journals and conferences to his credit. He is also a member of the editorial board of several international peer-reviewed journals and has been active Technical Program Committee member of several international and national conferences conducted by renowned universities and academic institutions. His field of expertise is Big Data, ANN, Biometrics, RFID, Cloud Computing, Cryptography, and VANETS.

Citation: *Dr. Gagandeep Jagdev et.al (2017). Scrutinizing and Executing the Positive Aspects of Big Data in World of Sports via Apache Hadoop Framework, International Journal of Research Studies in Computer Science and Engineering (IJRSCSE), 4(4), pp.36-42, DOI: http://dx.doi.org/10.20431/2349-4859.0404005*