[1.]Shafqat Ul AHSAAN, [2.]Ashish Kumar MOURYA

BIG DATA ANALYTICS: CHALLENGES AND TECHNOLOGIES

[1,2.]Department of Computer Science & Engineering, Jamia Hamdard, New Delhi, INDIA

Abstract: We are living in the era of Big Data. Nowadays, a huge volume of data that is complex too, is being generated from a variety of sources like Social Media, mobile phones, sensor networks, meteorological data, GPS data, healthcare data and surveillance data from drones and local cameras. These data generated from multiple sources if only stored in databases without any analysis then it is nothing other than garbage. To get value from this data, IT companies, research institutions, healthcare systems, data scientists are in a race to find the valuable patterns from this data to enhance decision making. The big challenge that landed with the existence of Big Data is: how to handle and manipulate such gigantic data. To process Big Data, the development of advanced and innovative tools and techniques is the need of the hour. The process of development in the IT infrastructure to an advanced level is itself a big challenge for IT industries. The challenge in the present scenario of Big Data is how to handle and manipulate an immense quantity of data that has to be firmly produced via the internet. This paper mainly focuses on the characteristics of Big Data and gives insights into a brief clarification of challenges related to Big Data and some analytical methods such as Hadoop and MapReduce.
Keywords: Big Data, Data Analytics, Hadoop, MapReduce, Data Mining

1. INTRODUCTION

The sudden increase of information that is being generated online by means of social media, internet, and worldwide communications has increasingly rendered data-driven learning. A new study revealed that over 4 million queries are being received by Google every minute, e-mails' sent by users reaches the limit of 200 million messages, 72 hours of videos are uploaded by YouTube users, 2 million chunks of content are shared over Facebook, and 277,000 Tweets are generated every minute on Twitter, Whatsapp users share 3,47,222 photos, Instagram users post 2,16,000 new photos every minute [1-2]. The present age is the age of Big Data, where data is growing on a large scale than ever before. According to the Computer World, 70% to 80% of data is considered to be in the unstructured form in organizations [3]. The data, which derives from social media, form 80% of the data globally and report for 90% of Big Data. As stated by the International Data Corporations (IDC) annual digital universe study [4], the data are being produced too rapidly and by the estimation of 2020, it would touch the range of 44 zettabytes which would be ten times larger than it was in 2013[5].

Today persistent and unified world, indeed, make people at the edge of an uninterrupted sensing process, where a mammoth quantity of data is produced and captured every minute. As per the study [6], each human is creating data through various sources on a large scale under the range of over 6 MB per minute, a sum of 1.7 million billion bytes of data. Also, the observance, authority, and Oversight commission declared that the size of data reaches twofold after every 18- 24 months for the majority of business organizations, whereas 90% of data have been produced in the last few years [7].

There are a vast number of domains belonging to dozens of organizations that have come to the point that they have to upgrade their technology infrastructure to handle the volume of data that have now touched the scale of Petabytes to remain in the competition. To handle Big Data, IT companies suggest their customers utilize the power of a mixture of technologies from NoSQL (NotonlySQL) databases like HBase, Paxata and Hadoop. However, Big Data architectures encounter an obstruction to adopt the analytics framework in terms of the level of complexity and standardization [8]. The shortage of skilled manpower (data scientists, developers, architects) and technical architecture is next level barrier in the field of Big Data analytics. It is required to focus on the requirements to cherish the overall potential of Big Data: 1) combining multiple expertise hands and system architectures and 2) Increasing the data manipulating and usability power of Big Data technologies.

Big Data can be elucidated on the basis of "3V" model proposed by Gartner as Big Data is a compilation of voluminous, frequently generated and multiple forms of data that require cost-effective, advanced tools in order to process the data for better understanding and decision making [9]. Big Data consists of both unstructured and structured data. Big Data is a complex, unstructured data that requires a series of advanced and unique techniques for its storage, processing, and analysis in order to transform it (data) into value. The flow of paper is as; introduction about Big Data is given in Section 1; characteristics of Big Data are covered in Section 2; Section 3 sheds light on challenges of Big Data, Section 4 mainly focused on core technologies for processing Big Data and finally, the conclusion is drawn in section 5.

2. CHARACTERISTICS OF BIG DATA

One perspective, accepted by Gartner's Doug Laney defined Big Data as composed of 3 Vs': Velocity, Variety, and Volume. Accordingly, IDC characterized it: "Big data techniques are related to the development of innovative technical

and architectural design in order to mine value from the massive quantity of multiple forms of data on a large scale, by allowing fast capture, finding, and/or analysis" [10]. This paper defines Big Data based on 8 Vs' as in Figure 1.

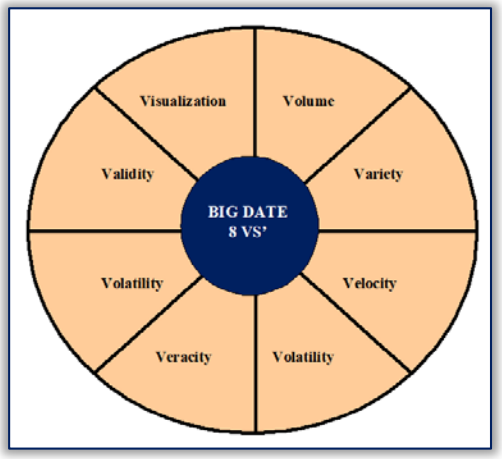

Figure 1. 8 Vs' of Big Data

- Volume. It is the infinite masses of data that are introduced each second.
- Velocity. How frequently the data is being generated by means of different sources.
- Variety. The multiple forms of data like music, pictures, text, financial transactions, videos, tweets, Facebook data; sensor data, etc. constitute a variety of data.
- Veracity. It is the intricacy of data which points to the certainty or quality of data. It means the meaningfulness of data.
- Validity. Validity and Veracity are not the same but have a similar concept. Validity means the accuracy of data for the intended usage. Veracity leads to validity if the data is properly understood, it means that we have to check properly and appropriately whether the dataset is valid for a particular application or not.
- Volatility. Volatility refers to the period for which we have to store the data. If volatility is not in place then a lot of storage space is wasted in storing data that is no more required, for instance, a commerce company keeps the purchase history of a customer for 1 year only as after 1 year the warranty on the purchased item expires so there is no reason to store such data.
- Value. It focuses on analytics and statistical methods, knowledge extraction and decision making.
- Visualization. Big Data visualization is the representation of data of nearly any category in a graphical layout that makes it easy to understand and interpret. Big Data visualization calls to mind the old saying: "a picture is worth a thousand words." That's because an image can often convey "what's going on", more quickly, more efficiently, and often more effectively than words. But it goes far beyond typical corporate graphs, histograms and pie charts to more complex representations like heat maps and fever charts, enabling decision makers to explore data sets to identify correlations or unexpected patterns.

3. CHALLENGES OF BIG DATA

As we know that Big Data is an emerging field of research and is still in the underdeveloped stage. It comes with a set of characteristics that make Big Data easy and simple to understand but unfortunately these characteristics have become a challenging factor. These characteristics make it a hindrance and a challenge to get useful information from. Researchers, data scientists, and organizations have to put every effort in order to cope up with these challenges and make Big Data a revolution to the field of science. Big Data exists with big challenges that are mainly categorized as data challenges, processing challenges, and management challenges.

- Data Challenges

i. Volume. The massive quantity of data that is derived every second constitutes the volume of Big Data [11]. There are multiple numbers of sources that play a key role in producing this vast portion of data like social media, surveillance cameras, sensor data, weather data, phone records, online transactions, etc. We are living in an age where data is generated in petabytes and zettabytes. This sudden boom in the production of data that is too large to store and analyze requires advanced tools and techniques that open the way for Big Data. To handle such voluminous data is really a big challenge for the data scientists.

ii. Velocity. Velocity is defined by how rapidly the new data is being generated. As we see how messages on social media go viral within no time, millions of photos are being uploaded by Facebook users each and every second, it takes milliseconds for the business systems to analyze social networking websites to gather message that set off the verdict to purchase or sell shares[12]. Big Data streaming processing method makes it possible to examine the data while it is emanated, in need of ever storing it into the database.

iii. Variety. Variety focuses on different forms of data like music, pictures, text, e-mails, medical records and images, weather records and log files, etc. generated from multiple sources. This means that the data produced belongs to different categories consisting of raw, unstructured, structured and semi-structured data which looks very difficult to deal with.

iv. Veracity. Veracity denotes the meaningfulness or value of data.

v. Value. Value focuses on the analytics and statistical methods, knowledge extraction and decision making [12]. The data that is generated and it is not analyzed and processed then it is nothing other than garbage.

– Processing Challenges

Data processing is common in every section of an organization. As we know, today data is derived through various sources like Facebook, Twitter, Whatsapp, mobile phones, sensors, surveillance cameras, etc. The more data we collect, the more accurate result will be available for optimization. To process this data generated on a large scale, new technologies, advanced algorithms and a variety of different approaches are required [12-13]. Processing of data comprises data gathering, fixing resemblance found in distinct sources, refinement of data to a type adequate for the scrutiny and output characterization. Let's have a look at various processing challenges of Big Data.

a) Heterogeneity. With the perspective of Big Data heterogeneity refers to a variety of data. There are infinite sources that produce multidimensional data on a large scale which constitutes Big Data. The data from different origin consists of dissimilar forms. Heterogeneity in Big Data is a requirement to tackle and deal with multiple forms of data concurrently [18].
b) Timeliness. Timeliness refers to how swiftly the data is being handled and analyzed. As the magnitude of the data that is set to be refined, get larger, undoubtedly will get extra time to evaluate. However, in some cases, the outcome of the evaluation is needed without delay. For example, if we have a transaction, and this transaction is made by a fraudulent, it must be identified before the transaction is over by blocking it from occurring. So here we have to build up unfinished consequences in advance so that a minute number of additional calculations with new data can be used to resolve the issue immediately [1-2].
c) Complexity. Complexity refers to the rigidness of data [4]. It mainly focuses on unstructured data. According to Zikopoulos and Eaton [8], Multidimensional Data can be classified into three types, namely, semi-structured, structured and unstructured. Structured data are homogeneous and lengths of data are defined in advance and are produced by automatic generators like computers or sensors without user communication. Unstructured data is complex such as client reviews, photos, and other multimedia. The unpredictable growth of the internet depicts that the size and range of Big Data keep on growing.
d) Scalability. As the data is generated at a large scale and is growing rapidly day-by-day. This rapid increase in the data rate can be mitigated by improving the processor speed. However, the quantity of data expands at a faster rate than computing resources and CPU speeds [13]. Taking all these parameters into consideration, we have to build such a scalable system with persistent storage and high processing speeds where a single node can share many hardware resources. In a broader sense, we can say scalability provides fault tolerance to the Big Data platform.
e) Accuracy. Accuracy refers that data is accurate and is obtained from limited sources. When we gather data from authentic sources and is extracted out from such data analysis, the results are more accurate [14].

– Management Challenges

The existing technologies of data management systems are incapable to gratify the requirements of Big Data, and the pace at which data storage space is increasing is far less than that of data generated, thus reconstruction of information framework in need of an hour. Besides, the existing algorithms do not fulfill the need to store the data efficiently that is directly acquired from different sources because of the heterogeneity of the data. We have pointed out a few important management challenges concerned with Big Data as below:

i. Security. As Big Data exceeds the data sources it can use, there is a need to verify the trust and worthiness of each source and therefore approaches should be used to find malevolent inserted data [15]. Information security is fetching to be a critical Big Data concern where a huge quantity of data will be linked scrutinized and obtained for significant patterns.
ii. Data sharing. We exist in the world of Big Data. Everything is being shared on social networking websites. Information about almost everything and anything can be collected by means of a single click on "Google". Each and every individual and corporation has an immense amount of information at their disposal that can be used for their purpose and requirement. Every kind of content is available and is one click away only when everyone shares it. But there must be differentiation of what is private and what can be shared [2], [16].
iii. Failure handling. To build a 100% reliable system is no way a straightforward task. Systems can be put up to such a degree that the possibility of breakdown must fall in the acceptable threshold. When a process is started, its processing involves multiple set-ups of nodes and the total working out process becomes unwieldy. Maintaining control points and setting up the threshold level for process resume subject to failure is a big challenge. Therefore, in Big Data this is called fault tolerance [17].

4. BIG DATA PROCESSING TECHNOLOGIES

– Hadoop Distributed File System

With the rise of Big Data, accessing Big Data was a big challenge for the researchers that are stored on a cluster in a distributed manner; the common solution that has figured out to access Big Data is a cluster file system. The cluster file system provides access to data files that can be stored anywhere on a cluster. One of the most popular cluster file systems

is the Hadoop Distributed File System (HDFS). Hadoop Distributed File System as is designed to provide the facility to store voluminous data in a large scale cluster and it works on master-slave architecture as depicted in figure 2.

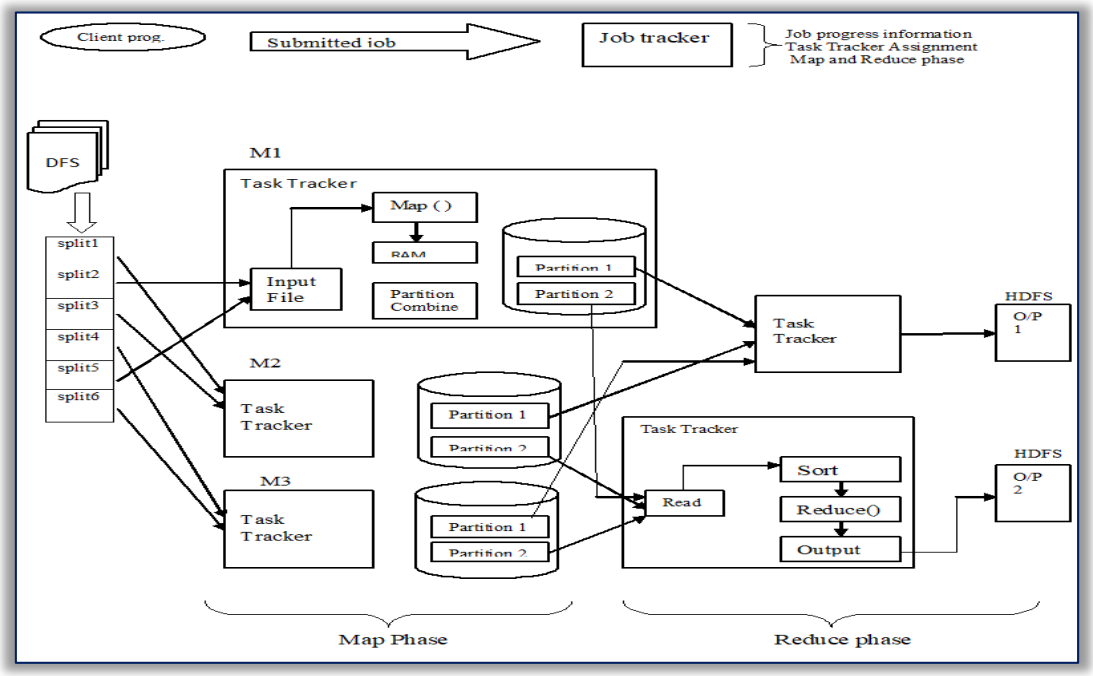

Figure 2. Hadoop Distributed File System (HDFS) architecture

It logically divides the file system into metadata and application data. In master-slave architecture, there are two types of computers called NameNode/ job tracker or master node and DataNode/ task tracker or slave node. The NameNode stores the metadata and another type of node called DataNodestores the application data. When a large data file is about to be stored, it splits into multiple blocks and these blocks are copied and stored among several DataNodes. In a Hadoop cluster, all the nodes are full-fledged connected in order to communicate with each other and the communication among different nodes in a cluster occurs with the help of TCP based protocols. The namespace of the file system is maintained by NameNode and the mapping of file blocks is maintained by DataNodes. When a file has to be read, NameNode is communicated to get the location where the data blocks are stored and then data blocks are read and viewed from the closest DataNode [18-20].

HDFS file system offers two striking features as compared to traditional file systems; it is highly fault-tolerant and can store data in petabytes. It supports high bandwidth and scalability.

– MapReduce

MapReduce is a programming model, used to refine for massive data files with the implementation of coordinated and disbursed algorithms on a cluster. MapReduce programming structure is sparked by the Map ( ) and the Reduce ( ) function.

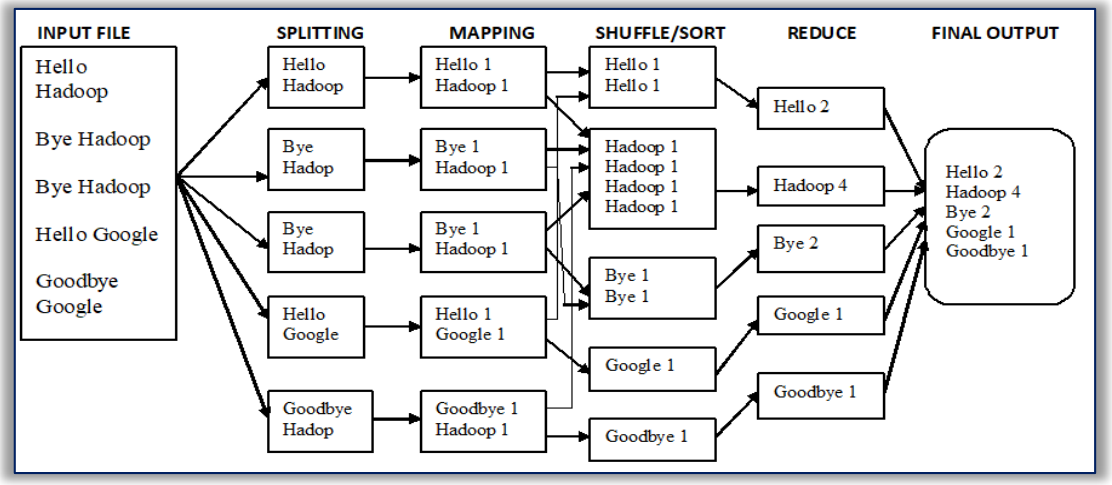

Figure 3. MapReduce Architecture

In Map ( ) step, the Master Node or the Name Node accepts the input file and partition it into minor sub-problems, these sub-problems are then assigned to Slave Nodes or Data Nodes. The Slave Nodes may further divide the problem into sub-subproblems. The Slave Node then handles these smaller problems and responds to the Master Node to which it is

connected. In the Reduce ( ) step, the Master Node receives the result and combines them together to turn out the final result to the original problem that it has to solve [19-20]. The working mechanism of MapReduce is depicted in Figure 3, few words are taken as input and then the various functions are performed on it, the final output is drawn to define how many numbers of times a specific word appears in the input file.

5. CONCLUSION

This paper entirely focuses on fundamental concepts about Big Data. In this paper, we have addressed and examined the number of challenges that comes in the way of Big Data. As we know that presently the data is generating at a very fast pace and will still continue to grow with every passing second. On the basis of research in the field of Big Data, the data will get doubled by the year 2020. To deal with such a gigantic data there is a need to rectify the challenges and hindrances that arise as a result of gigantic quantity data. The various challenges concerned with Big Data are briefly discussed. In addition, the researchers must focus on Big Data analysis that will boost up the economic position regarding commerce in the near future. Advanced tools and techniques are required to get value from Big Data such as Hadoop and MapReduce technologies. The Big Data V's model that defines the characteristics of Big Data is in point of fact the pitfalls we need to tackle and to resolve. To harvest the reimbursement of Big Data, we have to convert the challenges into power.

References:

[1] Jaseena, K.U. and David, J.M., 2014. Issues, challenges, and solutions: big data mining. CS & IT-CSCP, 4(13), pp.131-140.
[2] Lawal, Z., Zakari, R., Shuaibu, M. and Bala, A., 2016. A review: Issues and Challenges in Big Data from Analytic and Storage perspectives. International Journal of Engineering and Computer Science, 5(3), pp.15947-15961.
[3] Holzinger, A., Stocker, C., Ofner, B., Prohaska, G., Brabenetz, A. and Hofmann-Wellenhof, R., 2013. Combining HCI, natural language processing, and knowledge discovery-potential of IBM content analytics as an assistive technology in the biomedical field. In Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data (pp. 13-24). Springer, Berlin, Heidelberg.
[4] Landset, S., Khoshgoftaar, T.M., Richter, A.N. and Hasanin, T., 2015. A survey of open source tools for machine learning with big data in the Hadoop ecosystem. Journal of Big Data, 2(1), p.24.
[5] Jin, X., Wah, B.W., Cheng, X. and Wang, Y., 2015. Significance and challenges of big data research. Big Data Research, 2(2), pp.59-64.
[6] Damiani, E., Ardagna, C., Ceravolo, P. and Scarabottolo, N., 2017, September. Toward model-based big data-as-a-service: The toreador approach. In European Conference on Advances in Databases and Information Systems (pp. 3-9). Springer, Cham.
[7] Austin, D., 2012. eDiscovery Trends: CGOCs Information Lifecycle Governance Leader Reference Guide. 2012-05-03)[2013-06-10]. http://www. ediscoverydaily, com/2012/05/edis-covery-trends-cgocs-information-lifecycle-governance-leader-reference-guide, html.
[8] Ardagna, C.A., Ceravolo, P. and Damiani, E., 2016, December. Big data analytics as-a-service: Issues and challenges. In 2016 IEEE International Conference on Big Data (Big Data) (pp. 3638-3644). IEEE.
[9] Torre-Bastida, A.I., Del Ser, J., Laña, I., Ilardia, M., Bilbao, M.N. and Campos-Cordobés, S., 2018. Big Data for transportation and mobility: recent advances, trends and challenges. IET Intelligent Transport Systems, 12(8), pp.742-755.
[10] Ahmed, V., Tezel, A., Aziz, Z. and Sibley, M., 2017. The future of big data in facilities management: opportunities and challenges. Facilities, 35(13/14), pp.725-745.
[11] Kaur, H. and Wasan, S.K., 2006. Empirical study on applications of data mining techniques in healthcare. Journal of Computer science, 2(2), pp.194-200.
[12] Espinosa, J.A., Kaisler, S., Armour, F. and Money, W., 2019, January. Big Data Redux: New Issues and Challenges Moving Forward. In Proceedings of the 52nd Hawaii International Conference on System Sciences.
[13] Volk, M., Pohl, M. and Turowski, K., 2018. Classifying Big Data Technologies-An Ontology-based Approach.
[14] Ardagna, C.A., Ceravolo, P. and Damiani, E., 2016, December. Big data analytics as-a-service: Issues and challenges. In 2016 IEEE International Conference on Big Data (Big Data) (pp. 3638-3644). IEEE.
[15] Santos, A.F.C., Teles, Í.P., Siqueira, O.M.P. and de Oliveira, A.A., 2018. Big Data: A Systematic Review. In Information Technology-New Generations (pp. 501-506). Springer, Cham.
[16] Al-Khasawneh, M.A., Shamsuddin, S.M., Hasan, S. and Bakar, A.A., 2018, July. MapReduce a Comprehensive Review. In 2018 International Conference on Smart Computing and Electronic Enterprise (ICSCEE) (pp. 1-6). IEEE.
[17] Oussous, A., Benjelloun, F.Z., Lahcen, A.A. and Belfkih, S., 2018. Big Data technologies: A survey. Journal of King Saud University-Computer and Information Sciences, 30(4), pp.431-448.
[18] Gan, W., Lin, J.C.W., Fournier-Viger, P., Chao, H.C. and Yu, P.S., 2018. A survey of parallel sequential pattern mining. arXiv preprint arXiv:1805.10515.
[19] Talan, P.P., Sharma, K.U., Nawade, P.P. and Talan, K.P., 2019. An Overview of Hadoop MapReduce, Spark, and Scalable Graph Processing Architecture. In Recent Developments in Machine Learning and Data Analytics (pp. 35-42). Springer, Singapore.
[20] Bawankule, K., Singh, A.K. and Dewaang, R.K., 2019. Analysis of Task Scheduling in Hadoop MapReduce Framework. Australian Journal of Wireless Technologies, Mobility and Security e-ISSN 2200-1883, 1(1), pp.41-47.

ISSN 1584 - 2665 (printed version); ISSN 2601 - 2332 (online); ISSN-L 1584 - 2665