I. J. Education and Management Engineering, **2020, 4, 8-17**
Published Online August 2020 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijeme.2020.04.02

A Roadmap Towards Big Data Opportunities, Emerging Issues and Hadoop as a Solution

Rida Qayyum
Deaprtment of Computer Science, Government College Women University Sialkot, 51040, Pakistan
E-mail: ridaqayyum6@gmail.com

Received:24 May 2020; Accepted: 06 June 2020; Published: 08 August 2020

Abstract: The concept of Big Data become extensively popular for their vast usage in emerging technologies. Despite being complex and dynamic, big data environment has been generating the colossal amount of data which is impossible to handle from traditional data processing applications. Nowadays, the Internet of things (IoT) and social media platforms like, Facebook, Instagram, Twitter, WhatsApp, LinkedIn, and YouTube generating data in various formats. Therefore, this promotes a drastic need for technology to store and process this tremendous volume of data. This research outlines the fundamental literature required to understand the concept of big data including its nature, definitions, types, and characteristics. Additionally, the primary focus of the current study is to deal with two fundamental issues; storing an enormous amount of data and fast data processing. Leading to objectives, the paper presents Hadoop as a solution to address the problem and discussed the Hadoop Distributed File System (HDFS) and MapReduce programming framework for storage and processing in Big Data efficiently. Future research directions in this field determined based on opportunities and several emerging issues in Big Data domination. These research directions facilitate the exploration of the domain and the development of optimal solutions to address Big Data storage and processing problems. Moreover, this study contributes to the existing body of knowledge by comprehensively addressing the opportunities and emerging issues of Big Data.

Index Terms: Big Data, Internet of things (IoT), Social Media, Big Data Analytics, Hadoop, HDFS, MapReduce, YARN.

1. Introduction

With the evolution of technology, landline phones evolved with smartphones and tablets i.e. Google's Android and Apple iOS for making our life smarter. Apart from that, we have also used a bulky desktop for processing MBs of data. For repository, we used floppies, then hard disk for storing TBs of data and now storing data in the cloud for taking the edge off [1]. With this enhancement of existing technologies, data has been generated on a large scale. When we notice how much data generated through smartphones, we come to know that one video sends through WhatsApp or any other massager app generates data with every action we do [2]. Thus, the relational database is unable to handle this format of data. Despite the fact that the size of data also grows epidemically.

Self-driving cars came up having sensors that record every minute detail i.e. side of the obstacle, distance from the obstacle and many more in order to decide how to react. Meanwhile, data is generated for each kilometer that the user drives on that car [3]. Here, self-driving car is an example of Internet of things (IoT) that makes a device smarter by connecting the physical device with the internet [4]. For instance, smart air conditioner, this device monitors the temperature of the human body and the outside temperature. Correspondingly, determine what should be the temperature of the room. Now, for this, it first assembles the data from the internet through a sensor in order to monitor human body temperature. Meanwhile generating a huge amount of data.

Social media is one of the most significant elements in the advancement of big data [5]. Nowadays, everyone is using Facebook, Instagram, Twitter, YouTube and many other social media sites having millions of data such as personal detail, apart from that, each picture user like or react also generate data, liking Facebook page also generate data, sharing videos on Facebook result in generating a huge amount of data [6]. The most demanding aspect is that generated data do not exist in a structured manner but present in various formats and simultaneously, it is huge in volume and size. Table 1 describes the rapid production of data in various organizations further.

Table 1. Rapid growth of unstructured data

Source	Production
YouTube [7]	(i) Users upload 100 hours of new videos per minute (ii) Each month, more than 1 billion unique users access YouTube (iii) Over 6 billion hours of video are watched each month, which corresponds to almost an hour for every person on Earth.
Facebook [8]	(i) Every minute, 34,722 Likes are registered (ii) 100 terabytes (TB) of data are uploaded daily (iii) Currently, the site has 1.4 billion users (iv) The site has been translated into 70 languages
Twitter [9]	(i) The site has over 645 million users (ii) The site generates 175 million tweets per day
Foursquare [10]	(i) This site is used by 45 million people worldwide (ii) This site gets over 5 billion check-ins per day (iii) Every minute, 571 new websites are launched
Google+ [11]	1 billion accounts have been created
Google [12]	The site gets over 2 million search queries per minute Every day, 25 petabytes (PB) are processed
Tumblr [12]	Blog owners publish 27,000 new posts per minute
Instagram [12]	Users share 40 million photos per day
LinkedIn [12]	2.1 million groups have been created

Big data [13] is defined as a group of data records that is very large and complex which becomes hard to process using traditional data processing applications or on-hand database system tools [14]. When the traditional system was invented in the beginning, we never anticipated dealing with such a volume of data [15]. Current technologies not only increased the amount of data, but it also showed us data is actually getting generated in various formats i.e. data generated with video and images is unstructured.

The main purpose of this study is to presents: (a) A comprehensive survey of Big Data characteristics [16] (b) A discussion on types of Big Data to understand the concept of this domain [17] (c) The development of a new opportunities coming with Big Data. (d) Highlighting the issues associated with Big Data.

Moving towards the goal of attaining the storage and processing of the tremendous volume of data, among all issues, the objective of the current study is to deal with two fundamental issues; storing the colossal amount of data and fast data processing [18]. These problems have a direct relation to the number of resources used. Increased resources consequently demand the storage of data present in various formats and fastly data processing as well. In order to achieve the research objective, the Hadoop cluster has been considered as a solution to address these issues. The huge amount of generated data stored by Hadoop Distributed File System (HDFS) Architecture and processing is done by MapReduce while YARN facilitates in managing the resources in a very magnificent way. Therefore, the problem considered in the current study is how to store and process a huge amount of data using the Hadoop cluster. This work contribute to the researchers to more involve in big data technology.

The rest of this paper is organized as follows. In Section 2, we have discussed the big data characteristics. In Section 3, types of big data have been described in detail. Section 4 presented big data as an opportunity. We have discussed the emerging issues of big data in Section 5. Section 6 gives Hadoop as a solution to those issues that are associated with big data. Section 7 gives the discussion. Section 8 contains the conclusion of the conducted work.

2. Big Data Characteristics

The following equation was developed to further understand the concept of big data which is using the ideas of Kaisler, Armour, Espinosa, and Money [19], along with the fundamentals of discrete mathematics. However, below characteristics also referred to as 5 V's of big data.

$$Big\ Data = \{Data\ Volume, Data\ Variety, Data\ Velocity, Data\ Value, Data\ Veracity\} \quad (1)$$

Volume: This characteristic is nothing but the colossal amount of data that is being generated or the enormous quantity of data coming from various sources like social media, banking, and government sector, etc. By the year 2020, data assembled by the digital universe will enlarge from 4.4 ZB today to around 44 ZB, or 44 trillion GB.

Value: This characteristic refers to driving meaningful data from the entire collection of big data. Hereafter, we perform certain analytics on the collected data by making sure this analysis gives some value to data which further helps us in business to grow, provide certain insights not possible earlier.

Veracity: This feature explains inconsistencies and uncertainties which are present in the data. During the process, some data packages are bound to drop. Under an obligation to fill up these misplaced data by establish the mining again, proceed it, then it may come up with good insights if possible.

Velocity: The rate at which all varieties of data cumulate together gives velocity. This characteristic measures the rate at which data is being generated. As the grown amount of masses, the web application got increased on the internet and accessed via mobile, computers, and laptops generated a huge amount of data.

Variety: This feature deals with different formats of data including images, video, JSON files, and social media data coming from various sources. In big data, these three formats of data are structured, unstructured, multi-structured, and semi-structured data as explained in next section.

3. Types of Big Data

For the interpretation of big data, it is essential to understand the classification of data based on their behavior. Basically, big data is widely classified into different types as shown in Fig. 1

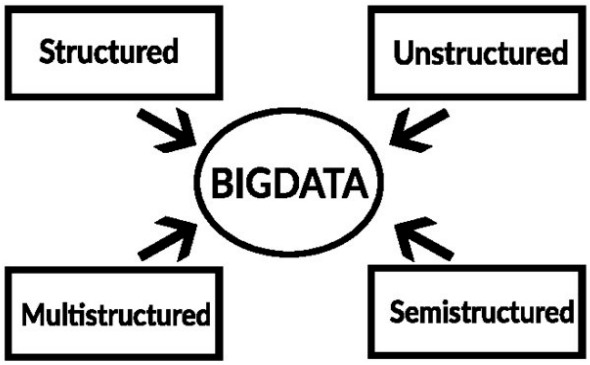

Fig. 1. Major Types of Big Data

3.1 Structured Data

Structured data are those type of data which are stored already in order. There are nearly 20% of the total existing data are structured data. All the data generated from sensors, weblogs are computationally structured. Human-generated structured data are those which are taken as information from humans such as their names, addresses, date of birth and gender, etc.

3.2 Unstructured Data

Unstructured data, on the other hand, is the one that has no clear format in storage. At least 80% of the data are unstructured. All satellite generated images, scientific data or images, and radar data are categorized as machine-generated unstructured data. There is various type of human-generated unstructured data such as images videos, social media data, PDFs, and text documents, etc.

3.3 Multi-structured Data

Multi-structure data is the type of data which have a different type of structure in it and can be derived from interactions between people and machines, such as web applications or social networks. A great example is web log data, which includes a combination of text and visual images along with structured data like form or transactional information.

3.4 Semi-structured Data

The semi-structured data incorporate both structured and unstructured data. This type of data sets cannot be stored in a traditional database structure, but it contains some organizational properties. Examples of semi-structured data are spreadsheet files, XML data, and JSON files, etc.

4. Big Data as an Opportunity

In this section, we have discussed the fields where we can use big data as a bone and there are certain unknown problems solved when dealing with big data, this bone referred to as big data analytics [20]. Today organizations are exploring large volume of data to discover those important facts which are never discovered before. As size of data is gradually increasing currently from few dozen terabytes to many petabytes of data in a single data set. So, big data analysis is required to gain valuable insights from these large and changing datasets. Hence, opportunities in big data have been discussed in table 2 as follows [21].

Table 2. Big Data Opportunities

Opportunities	Description
Cost effective storage system for dataset	Figure out how to store a data cost effectively without spending too much money on storage. With big data analytics, storage and management of the data become both reliable and feasible as compared to costly servers.
Fast and better decision making	Anticipate process to analyze information speedily and make decisions. Big data analytics helps business development, find more desirable and satisfactory insights from the data company have.
Improved products or services	Appraisal of customer need and fulfilment by analysing the organization collected data and making sure the customer preferences to generate recommendations for the company.
Next generation products	Revolutionizing different domains and products including healthcare, telecommunication, Smart yoga, Google's self-driving cars, and Netflix House of cards TV show.

It is significant to evaluate the data once accumulate to obtain insight. Hence, the issue is that these large data sets need to be stored and analyzed in order to extract value. The size, variety, and rapid change of such data require a new type of big data analytics, as well as different storage and analysis methods. Such sheer amounts of big data need to be properly analyzed so that pertaining information can be extracted. Primarily, there is four types of analyses which has been applied for all types of data [22] as depicted in Fig. 2.

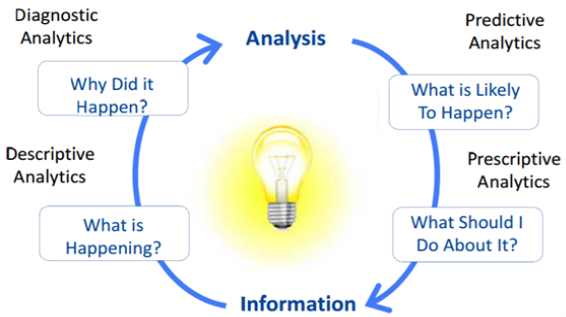

Fig. 2. Types of Big Data Analytics

4.1 Descriptive Analytics

By using data mining and data aggregation techniques, descriptive analysis provide insights in the past and then answer "What is happening now based on the incoming data?" Descriptive analysis is exactly what the name implies, it describe, comprises the raw data and make it something that is interpreted by humans. Google Analytics tool is the best example for descriptive analysis. A business get results from the internet using website by the devices which assist in comprehension what happened in the past and authenticate if a promotional campaign was victorious or not based on fundamental parameters like page views. Descriptive analysis is therefore, important source to determine what to do next. Let's have another example of Netflix, which is used in descriptive analysis to find the correlations among the different movies that a subscriber is watching and to upgrade the endorsement engine they utilize the customer data and significant sales.

4.2 Diagnostic Analytics

Diagnostic analytics is used to stimulate "Why something happens in the past?" It is categorized by the approaches like data mining, data discovery, drilled down and correlations. Moreover, it grasp the extensive aspects of the data to insight the main reason of the events. It is useful in stimulating what sort of factors and events conferred to a specific conclusion. Mostly, it uses probability, likelihood and the distribution of data for the analysis. For a Social Media

marketing campaign [23], you can use diagnostic analytics to determine the total followers, mentions, posts, fan, page views, pins and reviews, etc and analyses the failure and success rate of the campaign at an elementary level.

4.3 Predictive Analytics

Predictive analysis referred as using statistical framework, forecast strategy to insight the upcoming by responding "What could happen?" As the word suggest, it predict that what are the different future outcomes [24]. Basically, predictive analysis give actionable future directions to companies based on the data [25]. Through sensors and other machine created data, organizations can associate what a malfunction to be expected, then the firm preemptively arrange track and pre make restore to confront downtime and loses. For instance, Southwest Airline analyses sensor data on their planes to pinpoint repeated decorative figures that specify a possible malfunction, accordingly enabling the airline to the incumbent repairs before it schedule.

4.4 Prescriptive Analytics

Prescriptive analysis exert optimization and simulation algorithms to give direction on the achievable consequences and reply the query "What should be do?" Consequently, it permit the end user to advice numerous distinct feasible measures and then advise them about the solution. In a nutshell, prescriptive analysis is all about providing guidance. It utilize conjunction of tool and techniques such as algorithms, machine learning, business rules and computational modeling mechanism. Consequently, these approaches could applied against various distinct dataset including, real time data field, historical ad transactional data and big data. Google's self-driving car [26] is ideal illustration of prescriptive analytics. It analyses the surroundings and determine the direction to grasp depending on data. It choose whether to decrease or increase the speed to use another lane or not, to take a long cut to circumvent traffic or prefer ting roads etc.

5. Emerging Issues of Big Data

In this section, we have discussed the emerging issues associated with big data. So, the first problem is storing colossal amount of data. From past 2 years, the data originate has been beyond the expectation. By 2020, total digital data will enlarge to 44 ZB approximately and almost 1.7 MB of latest information will be generated time to time for every person. Thus, storing this huge data in traditional system is not possible and reason is that storage is limited to one system [27]. For instance, a company have server limit of 10 terabytes, but the company growing fast and data is exponentially increasing. Investing in huge servers is not a cost effective solution. A distributed file server is better solution for this huge data, this will saved the money.

Another problem is processing data having complex structure. Since, the data is not only huge but it is present in various formats as well like structured, unstructured, multi-structured, and semi-structured. It is essential to ensure the interconnected system is present to store this variety of data generated from various sources. Next problem focuses on fast data processing. Bringing huge amount of data to computation unit become a bottleneck. The data is growing at much faster rate than that of disk read. As hard disk capacity increases but disk transfer performance is not increasing at that rate as shown in Fig. 3.

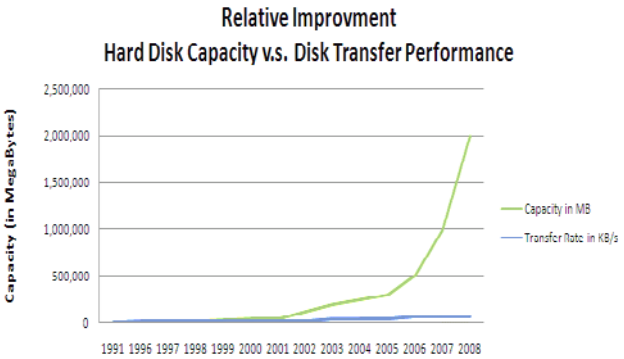

Fig. 3. Relative Improvement during data processing

6. Hadoop as a Solution

In this section, we have presented the Hadoop [28] as a solution to those issues that are associated with big data [29]. Hadoop is a framework that permits us to store and process large data sets in parallel and distributed manner [30]. It has the capability to process huge amounts of data mainly by allocating all the partitioned data sets to different nodes and each node solves a subpart of the problem and then combines all subparts to obtain the final result. Hadoop is a combination of Hadoop kernel, MapReduce, and Hadoop Distributed File System (HDFS). Basically, it has two parts, one is HDFS

(Storage) that allows parallel processing of the data in distributed file system [31] and other one is MapReduce (Processing) which enable to throw away any kind of data across the cluster through parallel and distributed processing [32]. There are two types of nodes in Hadoop: Master/Slave Architecture. The master node collects the input and distributes it to all the slave nodes in the map step. After this, the master node captures all the sub results and combines them to form the final result in the reduce step. The master node in the MapReduce framework is called the job tracker and all the slave nodes are called task trackers. The master node is responsible for job scheduling, fault tolerance, and distribution of subtasks to all the slaves. Moreover, we have discussed the Hadoop: Mater/Slave Architecture with a scenario. In a company, there is a project manager (Master Node) who handles a team of three employees (Slave Node). Whatever job a project manager gets from their clients, he distributes it across their team members and tracks the report on how the work is going on from time to time and it is the responsibility of each employee to complete the task. Here, exactly how Hadoop controls big data by practice the Master/Slave Architecture as depicted in Fig. 4.

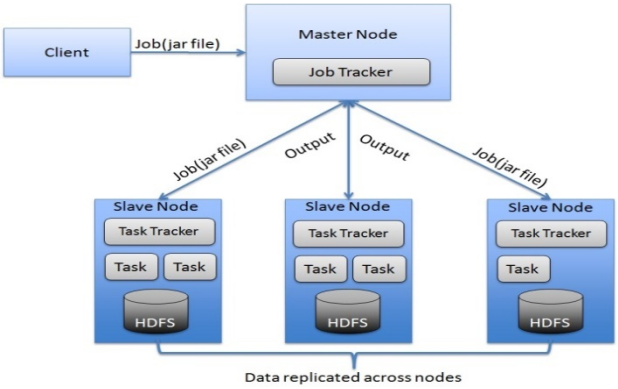

Fig. 4. Hadoop: Master/Slave Architecture

6.1 Hadoop Distributed File System (HDFS)

In order to solve the storage problem of big data we have HDFS [33]. All the big amount of data that we are dumping, it can distributing over different machines and these machines are interconnected on which data are distributed called Hadoop cluster. As depicted in Fig. 5, Hadoop Distributed File System (HDFS) architecture core components are:

NameNode: According to Hadoop: Master/Slave architecture, this node is refer as master node. It maintain and supervise all the distinct data nodes which are slave nodes in particular just like a project manager manages the team.

DataNode: Slave node are known as DataNode where the actual data stored in blocks. It is responsible for managing all data across data blocks. These nodes report the master node about the work progress by sending the signals known as heartbeats.

Secondary NameNode: This component is just a backup for the NameNode. It is the metadata, containing all the modifications took place across the Hadoop Cluster or HDFS namespace maintain by fsimage and editlog file.

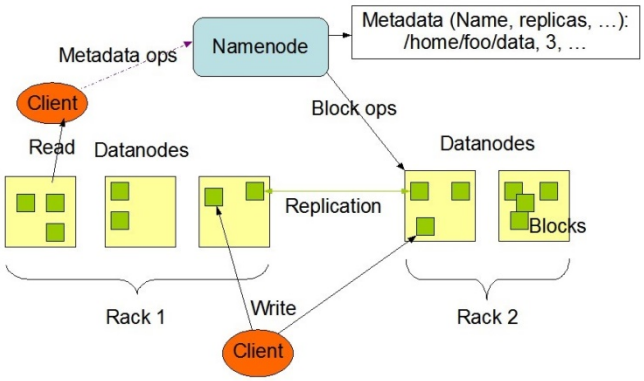

Fig. 5. Hadoop Distributed File System (HDFS) Architecture

6.2 MapReduce

In order to process big data, we have MapReduce programming framework as shown in Fig. 6. This framework permit the edge of utilize distributed framework to process enormous datasets [34]. The programming unit of Hadoop allows parallel and distributed processing of large data sets that is laying across the Hadoop cluster. Hence, MapReduce consist of two distinct tasks i.e. Map tasks and Reduce tasks. Every machine in Hadoop cluster processes the data which is known as Map. Finally, when the intermediary outcomes are combine to provide the final output this is called Reduce and

hence, collectively refer as MapReduce [35]. Table 3 introduces MapReduce tasks in job processing step by step.

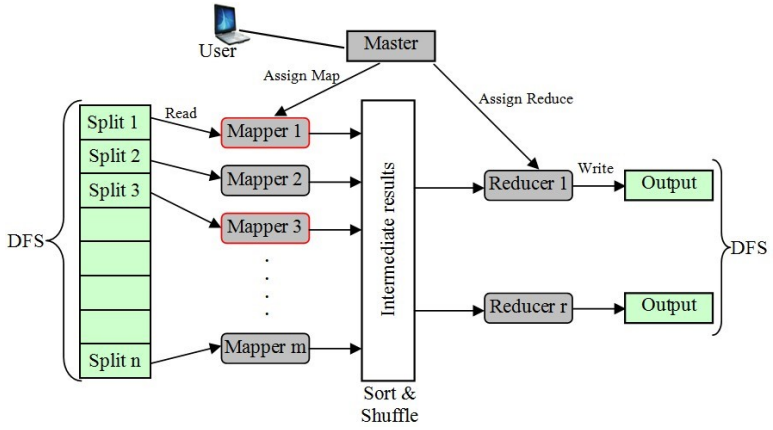

Fig. 6. MapReduce Programming Framework

Table 3. MapReduce Tasks

Steps	Tasks
(1) Input	(i) Data are loaded into HDFS in blocks an distributed to data nodes (ii) Blocks are replicated in case of failures. (iii) The name node tracks the blocks and data nodes
(2) Job	Submits the job and its details to the Job Tracker
(3) Job initialization	(i) The Job Tracker interacts with the Task Tracker on each data node (ii) All tasks are scheduled
(4) Mapping	(i) The Mapper processes the data blocks (ii) Key value pairs are listed
(5) Sorting	The Mapper sorts the list of key value pairs
(6) Shuffling	(i) The mapped output is transferred to the Reducers (ii) Values are rearranged in a sorted format
(7) Reduction	Reducers merge the list of key value pairs to generate the final result
(8) Result	(i) Values are stored in HDFS (ii) Results are replicated according to the configuration (iii) Clients read the results from the HDFS

6.3 YARN

In order to manage resources we have YARN (Yet Another Resource Negotiator) components including NodeManager, AppManager and container [36]. So, the Resource manager is the main node in the processing department. It receive processes requests like map produce jobs, then it passes the request to the NodeManager installed on every DataNode [37]. NodeManager and DataNode lies in a single machine which is responsible for AppManager and container. While AppManager monitors the MapReduce job is going fine and negotiate with Resource manager to ask for resources which might be needed to perform that particular MapReduce job. Container are combination of CPU and RAM. Fig. 7, depicts how MapReduce task takes place in YARN framework.

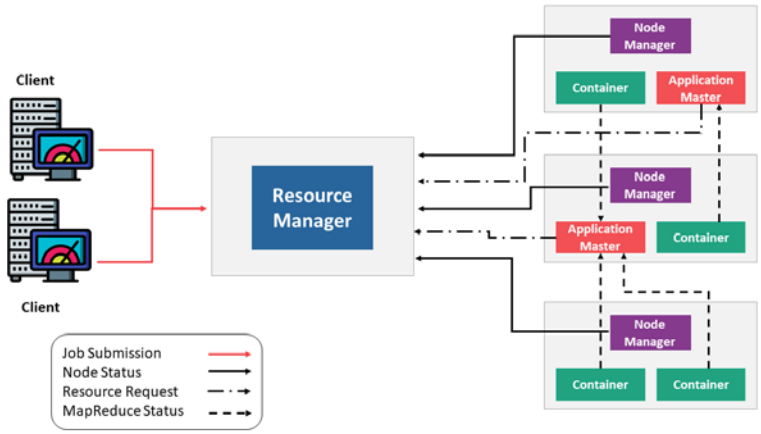

Fig. 7. Yet Another Resource Negotiator (YARN) Framework

7. Discussion

The primary object of the current research was to study a complete big data state-of-the-art opportunities, the problem with big data and solution to those issues which are highlighted. In order to understand the concept of big data, we have structured this paper as big data characteristics, big data types, the development of new opportunities coming with big data, emerging issues, and Hadoop as a solution. Nowadays, there are millions of smartphone users generating a massive amount of data through social networking, the Internet of things (IoT), and social media platforms. Additionally, storing and fast processing of big data becomes a challenge and it requires target investment to acquire big data technology advancement. The fundamental challenge mentioned in this paper is, data is not available in a structured mode and simultaneously it is colossal in size and volume demands fast and efficient processing.

The problem of storing data has been solved by Hadoop Distributed File System (HDFS) as it is a storage unit of Hadoop. Logically, being a single unit of big data, HDFS stores data across multiple systems in a distributed fashion by implementing the Hadoop: Master/Slave Architecture in which NameNode is a Master node and DataNode are Slaves while NameNode contains the metadata about the data that is stored in DataNode. Data blocks replicated based on task requirements. Basically, data has been stored in blocks and we specify the size of each block. If we have 512 MB of data and we have configured the HDFS such that it will create 128 Megabytes of data blocks. HDFS divides the data into four blocks each of size 128 MB storing across different DataNode. Here, we are using commandingly hardware and facing scaling challenges by focusing on horizontal scaling instead of vertical. For this, we always need to add some extra DataNode to your HDFS cluster for scaling the resources of data. Primarily, if we want to store 1 terabyte of data, we don't need 1 TB system. We can instead do it using multiple 2 GB of system or even less. Moreover, HDFS also address the storage issue of different data format. With HDFS, we can store all formats of data whether it is structured, unstructured, multi-structured, or semi-structured. We can dump all kinds of data we have in one place. In order to solve the fast processing of data, Hadoop MapReduce provides the solution. Having a master node and slave nodes. Data stored in the slave nodes. One way of processing the data, all the salve nodes send the data to the master node and the processing of data done at the master node. It will cause network congestion and input/output channel congestion and at the same time master node take a lot of time to process this huge amount of data. Actually, we need to send this process to data as all salve nodes already contain the data and perform processing as a slave itself. After that, the small chunks of the results comes out will be sent to NameNode. So, in that way, there will be no network overheads. Hence, Hadoop HDFS and MapReduce could be envisioned as a promising way of achieving the colossal amount of data storage and faster data processing.

8. Conclusion

We conclude that Big Data plays a vital role in current technologies and providing insights by generating a huge amount of data in various formats. Moreover, big data give various opportunities such as cost reduction, effective decision making, improved services or product, and next-generation products by analyzing the data with descriptive, diagnostic, predictive, and prescriptive analytics. As the volume of digital data tremendously increasing day by day, this aspect appears to be a giant issue for contemporary technologies to store and fast process this colossal amount data. The current study outline the basic literature necessary to comprehend big data and highlights these storage and processing problems and considered Hadoop as a solution. For storage, Hadoop Distributed File System (HDFS) permits parallel processing of the data in the distributed file system. On the other hand, MapReduce enables us to throw away any kind of data across the cluster through parallel and distributed processing. This research area is too big, we discuss some issues in this manner. For future work, further study on these issues can lead us to a better understanding of the big data domain. Whereas, Hadoop is

a technology that the future relies on, so the work can be done on the need to install Hadoop in a cloud server to manage big data and integrate with new frameworks so that Big Data will be widely accepted.

Acknowledgements

This work was performed under auspices of Department of Computer Science and Information Technology, Government College Women University Sialkot, Pakistan by Heir Lab-78. The Authors would like to thanks Dr. Muhammad Usman Ashraf for his continuous support, enthusiasm, insightful, and constructive suggestions throughout the research.

References

[1] Priyadarshini, S.B.B., BhusanBagjadab, A. and Mishra, B.K, "The role of IoT and big data in modern technological arena: A comprehensive study," Internet of Things and Big Data Analytics for Smart Generation, pp. 13-25, 2019.
[2] V.Mayer-Schonberger, K. Cukier, "Big Data: A Revolution That Will Transform How We Live Work and Think," Pub John Murray, pp. 256, 2013.
[3] Hong-Ning Dai, Hao Wang, Guangquan Xu, Jiafu Wan and Muhammad Imran, "Big data analytics for manufacturing internet of things: opportunities, challenges and enabling technologies", Enterprise Information Systems, 2019. DOI: 10.1080/17517575.2019.1633689
[4] Betty Jane J., Ganesh E.N, "Big Data and Internet of Things for Smart Data Analytics Using Machine Learning Techniques," Proceeding of the International Conference on Computer Networks, Big Data and IoT, vol. 49, 2019.
[5] Norjihan A. Ghani, Suraya H Ibrahim, Abaker T. Hashemb E. Ahmed, "Social media big data analytics: A survey," Elsevier Computers in Human Behavior, vol. 101, pp. 417-428, December 2019. DOI: 10.1016/j.chb.2018.08.039
[6] Jose L. J. Marquez Israel G. C. Jose, Luis L. C, Belen R. Mezcua, "Towards a big data framework for analyzing social media content", Elsevier International Journal of Information Management, vol. 44, pp. 1-12, February 2019, DOI: 10.1016/j.ijinfomgt.2018.09.003.
[7] YouTube, "YouTube statistics," 2014, http://www.youtube.com/ yt/press/statistics.html.
[8] Facebook, Facebook Statistics, 2014, http://www.statisticbrain .com/facebook-statistics/.
[9] Twitter, "Twitter statistics," 2014, http://www.statisticbrain .com/twitter-statistics/.
[10] Foursquare, "Foursquare statistics," 2014, https://foursquare .com/about.
[11] Jeff Bullas, "Social Media Facts and Statistics You Should Know in 2014," 2014, http://www.jeffbullas.com/2014/01/17/20-socialmedia-facts-and-statistics-you-should-know-in-2014/.
[12] Marcia, "Data on Big Data," 2012, http://marciaconner.com/ blog/data-on-big-data/.
[13] Younas, M., "Research challenges of big data," 2019.
[14] Al-Mekhlal, M. and Khwaja, A.A, "A Synthesis of Big Data Definition and Characteristics," IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC), pp. 314-322, August 2019.
[15] S. Madden, "From Databases to Big Data", IEEE Internet Computing, vol.16, no.3, pp. 4–6, 2012.
[16] El Alaoui, I., Gahi, Y. and Messoussi, R., "Full consideration of Big Data characteristics in sentiment analysis context," IEEE 4th International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), pp. 126-130, 2019.
[17] Khan, N., Naim, A., Hussain, M.R., Naveed, Q.N., Ahmad, N. and Qamar, S., "The 51 V's Of Big Data: Survey, Technologies, Characteristics, Opportunities, Issues and Challenges". Proceedings of the International Conference on Omni-Layer Intelligent Systems, pp. 19-24. May 2019.
[18] Aggarwal, A.K, "Opportunities and challenges of big data in public sector," Web Services: Concepts, Methodologies, Tools, and Applications, pp. 1749-1761, 2019.
[19] S. Kaisler, F. Armour, J. A. Espinosa, and W. Money, "Big Data: Issues and Challenges Moving Forward," in 46th Hawaii International Conference on System Sciences, 2013, pp. 995–1004.
[20] Dai, H.N., Wang, H., Xu, G., Wan, J. and Imran, M., "Big data analytics for manufacturing internet of things: opportunities, challenges and enabling technologies," Enterprise Information Systems, pp.1-25, 2019.
[21] Subramaniam, Anushree. "What Is Big Data Analytics | Big Data Analytics Tools and Trends | Edureka". Edureka, 2020, https://www.edureka.co/blog/big-data-analytics/.
[22] Dai, H.N., Wong, R.C.W., Wang, H., Zheng, Z. and Vasilakos, A.V, "Big data analytics for large-scale wireless networks: Challenges and opportunities," ACM Computing Surveys (CSUR), pp. 1-36., 2019.
[23] Shirdastian, H., Laroche, M. and Richard, M.O, "Using big data analytics to study brand authenticity sentiments: The case of Starbucks on Twitter," International Journal of Information Management, pp.291-307, 2019.
[24] Zhang, Z, "Predictive analytics in the era of big data: opportunities and challenges," Annals of Translational Medicine, 2020.
[25] C. Statchuk, M. Iles, F. Thomas, "Big data and analytics", in Proceedings of the 2013 Conference of the Center for Advanced Studies on Collaborative Research (CASCON 13), pp. 341–343. 2013.
[26] ur Rehman, M.H., Yaqoob, I., Salah, K., Imran, M., Jayaraman, P.P. and Perera, C., "The role of big data analytics in industrial Internet of Things," Future Generation Computer Systems, pp. 247-259, 2019
[27] Paliszkiewicz, J., "Management in the Era of Big Data: Issues and Challenges," 2020.
[28] Gupta, H.K. and Parveen, R., "Comparative Study of Big Data Frameworks," IEEE International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT), vol. 1, pp. 1-4, September 2019.
[29] Raj, A. and D'Souza, R., "A Review on Hadoop Eco System for Big Data," 2019.

[30] Hussain, T., Sanga, A. and Mongia, S, "Big Data Hadoop Tools and Technologies: A Review", 2019, Available at SSRN 3462554.
[31] Asim, M., McKinnel, D.R., Dehghantanha, A., Parizi, R.M., Hammoudeh, M. and Epiphaniou, G., "Big data forensics: Hadoop distributed file systems as a case study" Handbook of Big Data and IoT Security, pp. 179-210, 2019.
[32] Deshai, N., Sekhar, B.V.D.S., Venkataramana, S., Srinivas, K. and Varma, G.P.S, "Big Data Hadoop MapReduce Job Scheduling: A Short Survey," Information Systems Design and Intelligent Applications, pp. 349-365, 2019.
[33] Hu, F., Yang, C., Jiang, Y., Li, Y., Song, W., Duffy, D.Q., Schnase, J.L. and Lee, T., "A hierarchical indexing strategy for optimizing Apache Spark with HDFS to efficiently query big geospatial raster data," International Journal of Digital Earth, pp.410-428, 2020.
[34] Deshai, N., Sekhar, B.V.D.S., Venkataramana, S., Srinivas, K. and Varma, G.P.S., "Big Data Hadoop MapReduce Job Scheduling: A Short Survey," Information Systems Design and Intelligent Applications, pp. 349-365, 2019.
[35] Lev-Libfeld, A. and Margolin, A., "Fast Data: Moving beyond from Big Data's map-reduce," arXiv preprint arXiv:1906.10468
[36] Monu, M. and Pal, S., "A Review on Storage and Large-Scale Processing of Data-Sets Using Map Reduce, YARN, SPARK, AVRO, MongoDB. YARN, SPARK, AVRO, MongoDB," April 2019.
[37] Li, R., Yang, Q., Li, Y., Gu, X., Xiao, W. and Li, K, "HeteroYARN: a heterogeneous FPGA-accelerated architecture based on YARN," IEEE Transactions on Parallel and Distributed Systems. 2019.

Authors' Profile

Rida Qayyum (born September 17, 1996) is a graduate student of Bachelor of Science in Information Technology (BSIT), Department of Computer Science from Government College Women University, Sialkot. She attended a seminar on Cyber Secure Pakistan organized by PISA at the National Library of Pakistan Islamabad and also certified as Microsoft Office Specialist. She has many publications in international journals. Her research interest including Location Based Services (LBS) Systems, Network Security, Cloud Computing, Big Data, Deep Learning, and Parallel Computing.

How to cite this paper: Rida Qayyum. " A Roadmap Towards Big Data Opportunities, Emerging Issues and Hadoop as a Solution ", International Journal of Education and Management Engineering (IJEME), Vol.10, No.4, pp.8-17, 2020. DOI: 10.5815/ijeme.2020.04.02

