Journal of Advances and Scholarly Researches in Allied Education
Vol. 16, Issue No. 3, (Special Issue) March-2019, ISSN 2230-7540

Big Data Analysis and Its Security Challenges

Priyanka Gautam*

Professor of Computer and Information Science, CPJCHS, Narela

Abstract – "Big Data" provides futuristic techniques and mechanisms to store, distribute, capture, manage and examine petabyte or larger-sized datasets with high-velocity and different shapes. Big data can be structured, unstructured or semi-structured, resulting in inability of ordinary data management methods. Data is produced from various different sources and can arrive in the system at various rates. In order to action these large amounts of data in areasonable and efficient way, parallelism is used. Big Data is a data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it. Hadoop is the core platform for structuring Big Data, and solves the problem of making it useful for analytics purposes. This paper aims to analyze some of the different analytics methods and tools which can be applied to big data, as well as the opportunities provided by the application of big data analytics in Various decision domains.

Keywords: Big data, Data Mining, Analysis, Hadoop Framework, HDFS, map reduce, Hadoop Component

----------X----------

INTRODUCTION

1. Big data: immensely large datasets that is tough to deal with using Relational Databases Storage/Cost, Search/Performance, Analytics and Visualization. There's no particular method defined to determine whether the particular size of data comes under the category of big data or no and also data continues to change over time, most analysts and practitioners currently refer to data sets from 30-50 terabytes (10 12 or 1000 gigabytes per terabyte) to multiple petabytes (1015 or 1000 terabytes per petabyte) as big data. Every second, more and more data is being created and needs to be stored and analyzed in order to extract value. Furthermore, data has become cheaper to store, so organizations need to get as much value as possible from the huge amounts of stored data.

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, and pertaining information should be extracted.

2. 4 V's of Big Data:

i. Volume of data: The amount of data is known as volume. Volume of data stored in enterprise repositories have grown from megabytes and gigabytes to petabytes 40 Zetta bytes of data will be created by 2020 which is 300 times from2005.

ii. Variety of data: Different types of data and sources of data. Data variety exploded from structured and legacy data stored in enterprise repositories to unstructured, semi structured, audio, video, XML etc.

iii. Velocity of data: Velocity refers to the speed of data processing. For time-sensitive processes such as catching fraud, big data must be used as it streams into your enterprise in order to maximize its value.

iv. Veracity of data: Veracity means accuracy of data. Data is uncertain due to the inconsistency and in completeness. Veracity means anxiety or accuracy of data. Data is uncertain due to the inconsistency and in completeness. Veracity means anxiety or accuracy of data. Data is uncertain due to the inconsistency and in completeness.

3. Big Data Analytics: The term "Big Data" has recently been applied to datasets that grow so large that they become awkward to work with using traditional database management systems. They are data sets whose size is beyond the ability of commonly used software tools and storage systems to capture, store, manage, as well as process the data within a tolerable elapsed time (Kubick, 2012). Big data sizes are constantly increasing, currently ranging from a few dozen terabytes (TB) to many petabytes (PB) of data in a single data set. Consequently, some of the

difficulties related to big data include capture, storage, search, sharing, analytics, and visualizing. Today, enterprises are exploring large volumes of highly detailed data so as to discover facts they didn't know before .Hence, big data analytics is where advanced analytic techniques are applied on big data sets. Analytics based on large data samples reveals and leverages business change. However, the larger the set of data, the more difficult it becomes to manage (Russom, 2011). In this section, we will start by discussing the characteristics of big data, as well as its importance. Naturally, business benefit can commonly be derived from analyzing

larger and more complex data sets that require real time or near-real time capabilities; however, this leads to a need for new data architectures, analytical methods, and tools. Therefore the successive section will elaborate the big data analytics tools and methods, in particular, starting with the big data storage and management, then moving on to the big data analytic processing. It then concludes with some of the various big data analyses which have grown in usage with big data.

3.1 Characteristics of Big Data

Big data is data whose scale, distribution, diversity, and/or timeliness require the use of new technical architectures, analytics, and tools in order to enable insights that unlock new sources of business value. Three main features characterize big data: volume,

Variety, velocity, or the three V's. The volume of the data is its size, and how enormous it is. Velocity refers to the rate with which data is changing, or how often it is created. Finally, variety includes the different formats and types of data, as well as the different kinds of uses and ways of analyzing the data (EMC, 2012). Data volume is the primary attribute of big data. Big data can be quantified by size in TBs or PBs, as well as even the number of records, transactions, tables, or files. Additionally, one of the things that make big data really big is that it's coming from a greater variety of sources than ever before, including logs, clickstreams, and social media. Using these sources for analytics means that common structured data is now joined by unstructured data, such as text and human language, and semi-structured data, such as EXtensible Markup Language (XML) or Rich Site Summary (RSS) feeds. There's also data, which is hard to categorize since it comes from audio, video and other devices. Furthermore, multi-dimensional data can be drawn from a data warehouse to add historic context to big data. Thus, with big data, variety is just as big as volume. Moreover, big data can be described by its velocity or speed. This is basically the frequency of data generation or the frequency of data delivery. The leading edge of big data is streaming data, which is collected in real-time from the websites (Russom, 2011). Some researchers and organizations have discussed the addition of a fourth V, or veracity. Veracity focuses on the quality of the data. This characterizes big data quality as good, bad, or undefined due to data inconsistency, incompleteness, ambiguity, latency, deception, and approximations.

4. Challenges with Big Data Security

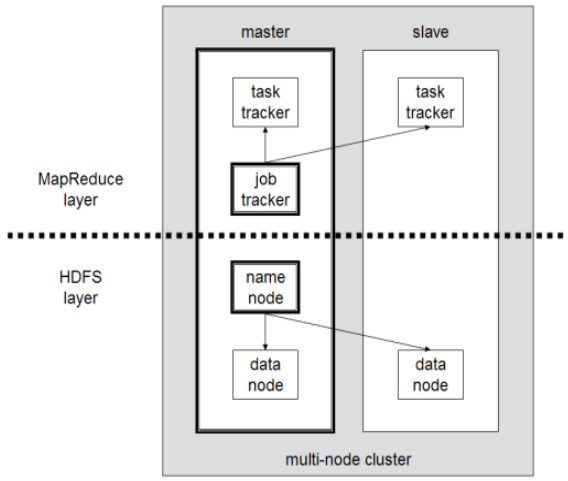
i) **Heterogeneity and Incompleteness:** When humans consume information, a great deal of heterogeneity is comfortably tolerated. In fact, the richness of natural language can provide valuable depth. However, machine analysis algorithms expect homogeneous data, and cannot understand nuance. In consequence, data must be carefully structured as a first step in (or prior to) data analysis. Computer systems work most efficiently if they can store multiple items that are all identical in size and structure.

ii) **Privacy:** Privacy of data is one bigger problem with big data. In some countries, there are tough laws about the data privacy, for example in USA there are tough law for health records, but for others it is less forceful. For example in social media we cannot get the private posts of users for sentiment analysis.

iii) **Scale:** As the name says Big Data is having huge size of data sets. Managing with large data sets is a big problem from decades. In previous years, this problem was solved by the processors getting faster but now data quantity is becoming large and processors are static. World is moving towards the Cloud technology, due to this shift data is generated in a very high rate. This high rate of increasing data is becoming a challenging problem to the data analysts. Hard disks are used to store the Data. They are slower I/O performance.

iv) **Human Collaborations:** In spite of the tremendous advances made in computational analysis, there remain many patterns that humans can easily detect but computer algorithms have a hard time finding. Ideally, analytics for Big Data will not be all computational rather it will be designed explicitly to have a human in the loop. The new sub-field of visual analytics is attempting to do this, at least with respect to the modelling and analysis phase in the pipeline.

5. Hadoop Architecture



Hadoop is an open source project hosted by Apache Software Foundation. It consists of many small sub projects which belong to the category of infrastructure for distributed computing. Hadoop mainly consists of:

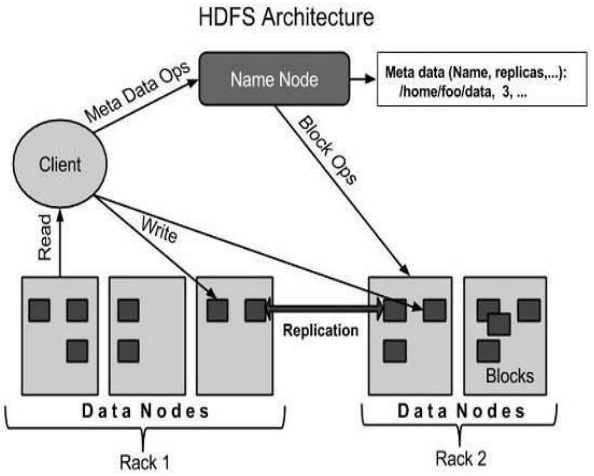
1. File System (The Hadoop File System)

2. Programming Paradigm (Map Reduce)

Hadoop was developed by Google's MapReduce that is a software framework where an application break down into various parts. The Current Apache Hadoop ecosystem consists of the Hadoop Kernel, MapReduce, HDFS and numbers of various components like Apache Hive, Base and Zookeeper. HDFS and MapReduce.

5. i) Hadoop Distributed File System

Hadoop develop with a distributed File System called HDFS, HDFS stands for Hadoop Distributed File System. The Hadoop Distributed File System is a versatile, clustered way to handling files in a big data environment. HDFS is not the final terminal for files. It is a kind of data service that offers a different set of capabilities required when data volumes and velocity are high. Because the data is written once and then read many times. HDFS is a good choice for supporting big data analysis.



HDFS follow the master slave and architecture.

a. Name Node

It is centrally placed node, which contains information about Hadoop file system . The main task of name node is that it records all the metadata & attributes and specific locations of files & data blocks in the data nodes. Name node acts as the master node as it stores all the information about the system and provides information which is newly added, modified and removed from data nodes.

b. Data Node

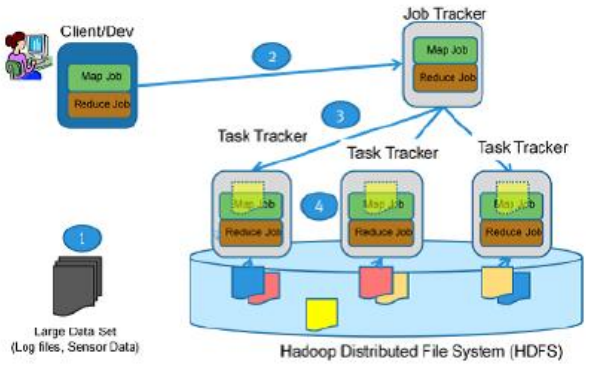
It works as slave node. Hadoop environment may contain more than one data nodes based on capacity and performance. A data node performs two main tasks storing a block in HDFS and acts as the platform for running jobs.

5. (ii) Map Reduce Framework

MapReduce is defined as a programming model for processing and generating large sets of data. There are two phases in MapReduce, the "Map" phase and the "Reduce" phase. The system splits the input data into multiple chunks, each of which is assigned a map task that can process the data in parallel. Each map task reads the input as a set of (key, value) pairs and produces a transformed set of (key, value) pairs as the output. The framework shuffles and sorts outputs of the map tasks, sending the intermediate (key, value) pairs to reduce task, which groups them into final results.

map – The function takes key/value pairs as input and generates an intermediate set of key/value pairs

Reduce – the function which merges all the intermediate values associated with the same intermediate key



6. Security issues and challenges:

Challenges Related to Characteristics of Big Data

- **Data volume:** storage is the very first issue that comes in as we thing about the volume. As data volume increases so the amount of space required to store data efficiently also increases. Not only that the huge volumes of data needs to be retrieved at a fast speed to extract results from them. Networking, bandwidth, cost of storing like in-house versus cloud storing are other areas to be looked after (Adams, 2010). With the increase in volume of data the value of data records tends to decrease in proportion to age, type, richness and quality (Asur & Huberman, 2010). The advent of social networking sites have led to production of data of the order of terabytes every day. Such volumes of data are difficult to be handled using existing traditional databases (Asur & Huberman, 2010).

- **Data velocity:** Computer systems are creating more and more data, both operational and analytical at increasing speeds and the number of consumers of that data are growing. People want all of the data and they want it as soon as possible leading to what is trending as high-velocity data. High velocity data can mean millions of rows of data per second. Traditional database systems are not capable enough of performing analytics on such volumes of data and that is constantly in motion. Data generated by both devices and actions of human beings like log files, website clickstream data like in E-commerce; twitter feeds can't be collected because the state of the art technology can't handle that data (Asur & Huberman, 2010).

- **Data variety:** Big data comes in many a form like messages, updates and images in social media sites, GPS signals from sensors and cell phones and a whole lot more. Many of these sources of big data are virtually new or rather as old as the networking sites themselves, like the information from social networks, Facebook, launched in 2004 and Twitter in 2006. Smart phones and other mobiles devices can be bracketed in the same category. As these devices are ubiquitous the traditional databases that store most corporate information until recently are found to be ill suited to these data. Much of these data are unstructured and unwieldy and noisy which requires rigorous technique for decision making based on the data. Better algorithms to analyze them are an issue too (Cohen, et. al., 2009).

- **Data value:** Data are stored by different organizations to gain insights from them and use them for analytics for business Intelligence. This storing produces a gap between the business leaders and the IT professionals. The business leaders are concerned with adding value to their business and obtaining profits from it. More the data more are the insights. This however doesn't go well with the IT professionals as they have to deal with the technicalities related to storing and processing the huge amounts of data.

CONCLUSION:

The paper describes the concept of Big Data along with 3 Vs, Volume, Velocity and variety of Big Data and also, paper discusses some of the security issues that come in when we work on big data. This paper discusses some basic concept of Architecture of big data Hadoop along with security problems. Hadoop which is an open source software used for processing of Big Data. In this research, we have examined the innovative topic of big data, which has recently gained lots of interest due to its perceived unprecedented opportunities and benefits. In the information era we are currently living in, voluminous varieties of high velocity data are being produced daily, and within them lay intrinsic details and patterns of hidden knowledge which should be extracted and utilized. Hence, big data analytics can be applied to leverage business change and enhance decision making, by applying advanced analytic techniques on big data, and revealing hidden insights and Valuable knowledge.

REFERENCE:

1. Adams, M.N. (2010). Perspectives on Data Mining. International Journal of Market Research 52(1), pp. 11–19.

2. Asur, S., Huberman, B.A. (2010). Predicting the Future with Social Media. In: ACM International Conference on Web

Intelligence and Intelligent Agent Technology, vol. 1, pp. 492–499.

3. Bakshi, K. (2012). Considerations for Big Data: Architecture and Approaches. In: Proceedings of the IEEE Aerospace Conference, pp. 1–7.

4. CEBR (2012). Data equity, unlocking the value of big data. in: SAS Reports, pp. 1–44.

5. Cohen, J., Dolan, B., Dunlap, M., Hellerstein, J.M., Welton, C.: MAD Skills (2009). New Analysis Practices for Big Data. Proceedings of the ACM VLDB Endowment 2(2), pp. 1481–1492.

6. Cuzzocrea, A., Song, I., Davis, K.C. (2011). Analytics over Large-Scale Multidimensional Data: The Big Data Revolution! In: Proceedings of the ACM International Workshop on Data Warehousing and OLAP, pp. 101–104.

7. Economist Intelligence Unit: The Deciding Factor: Big Data & Decision Making. In:Capgemini Reports, pp. 1–24 (2012).

8. Elgendy, N. (2013). Big Data Analytics in Support of the Decision Making Process. MSc Thesis, German University in Cairo, p. 164.

9. EMC (2012). Data Science and Big Data Analytics. In: EMC Education Services, pp. 1–508.

10. He, Y., Lee, R., Huai, Y., Shao, Z., Jain, N., Zhang, X., Xu, Z. (2011). RC File: A Fast and Space efficient Data Placement Structure in MapReduce-based Warehouse Systems. In: IEEE International Conference on Data Engineering (ICDE), pp. 1199–1208.

11. Herodotou, H., Lim, H., Luo, G., Borisov, N., Dong, L., Cetin, F.B., Babu, S. (2011). Starfish: A Self-tuning System for Big Data Analytics. In: Proceedings of the Conference on Innovative Data Systems Research, pp. 261–272.

12. Kubick, W.R. (2012). Big Data, Information and Meaning. In: Clinical Trial Insights, pp. 26–28.

13. Lee, R., Luo, T., Huai, Y., Wang, F., He, Y., Zhang, X. (2011). Ysmart: Yet Another SQL-to- MapReduce Translator. In: IEEE International Conference on Distributed Computing Systems (ICDCS), pp. 25–36.

14. Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A.H. (2011). Big Data: The Next Frontier for Innovation, Competition, and Productivity. In: McKinsey Global Institute Reports, pp. 1–156.

15. Mouthami, K., Devi, K.N., Bhaskaran, V.M. (2011). Sentiment Analysis and Classification Based on Textual Reviews. In: International Conference on Information Communication and Embedded Systems (ICICES), pp. 271–276.

16. Plattner, H., Zeier, A. (2011). In-Memory Data Management: An Inflection Point for Enterprise Applications. Springer, Heidelberg.

17. Russom, P. (2011). Big Data Analytics. In: TDWI Best Practices Report, pp. 1–40.

18. Sanchez, D., Martin-Bautista, M.J., Blanco, I., Torre, C. (2008). Text Knowledge Mining: An Alternative to Text Data Mining. In: IEEE International Conference on Data Mining Workshops, pp. 664–672.

19. Serrat, O. (2009). Social Network Analysis. Knowledge Network Solutions 28, 1–4.

20. Shen, Z., Wei, J., Sundaresan, N., Ma, K.L. (2012). Visual Analysis of Massive Web Session Data. In: Large Data Analysis and Visualization (LDAV), pp. 65–72.

Corresponding Author

Priyanka Gautam*

Professor of Computer and Information Science, CPJCHS, Narela

parrygautam5@gmail.com