ISSN 2348-1218 (print)
International Journal of Interdisciplinary Research and Innovations ISSN 2348-1226 (online)
Vol. 9, Issue 1, pp: (29-32), Month: January - March 2021, Available at: www.researchpublish.com

BIG DATA REVIEW BASED ON SECURITY CHALLENGES AND PRIVACY ISSUES

Yazeed Al Moaiad [1], Wafa Al-Haithami [2]

Al-Madinah International University, Malaysia

[1]yazeed.alsayed@mediu.edu.my

[2]wafaaalhithmy@yahoo.com

Abstract: **This paper is about the trend of gathering, storing, and managing high volume data sets known as big data. It will be gone to introduce the term and review some basics and characteristics of Big Data followed by security challenges and privacy issues. The researchers will also discuss the three versus (Velocity, Volume, and Variety) of big data with the addition of two recently added versus (Veracity and Value).**

Keywords: **Big Data, Hadoop, Apache, Velocity, Volume, Variety.**

I. INTRODUCTION

As we can see, we are turning to the age of speed in all aspects of life, which helps to save a large variety of data from different devices. Therefore, processing data by traditional methods becomes nearly impossible, which reduces the collection and analysis of data to produce specific insights and concepts for that data. If we want to process that data by using relational database engines, then it is not possible because the data may be disorganized, sensitive to time, and most importantly, it may be very large. Therefore, this data requires a different method of processing it in terms of capacity and approach, which has come to be called "big data"[1].

It is through knowing the big data reformer who is defined as a description of collecting and storing big data. This big data is either sequentially organized, or it is not randomly organized, but on the other hand, processing that big data helps in better decision-making and defining the strategic goals of the business. Also, increasing knowledge of a certain thing will increase reliability, gain new perspectives and insights, and help to better predict the decisions to be made later. Comparing data and the relationships between them helps in raising the efficiency of learning to make appropriate and smarter decisions.

Therefore, companies do not care about the size of big data, but the most important thing is what they will use this data, not after processing it, because in its initial stages it has no value until it is processed with the right methods and tools. So, the goal of this big data is to process it from the original pattern into a mature, usable idea, and hence the problem with this big data.

II. LECTURE REVIEW

The term "big data" is considered relatively new according to [2], as it has been known to professionals in this field to collect, store and analyze big data from a long time. Accordingly, the term gained momentum when industry analyst Doug Laney explained the definition of big data as a concept of three components: volume, variety, and velocity. Indeed, the author [3], added two other elements, namely value and honesty, during the past few years.

The term "big data" was discussed by the author [4], from a certain angle, as he said that this concept may have been around for quite some time, but there was and still is some ambiguity and confusion as to what exactly this term means. He pointed out that this concept is developing over time, so what it means should be reconsidered, as it remains the driving force towards continuous digital development, and what we are witnessing in the field of industrial, Internet of things and data science is the best evidence of that.

When talking about this term "big data", which may be relatively new as indicated by the author [2,3 and 4], the origin of big data may go back to the sixties and seventies of the current century, when the data was in a forced increase, and its importance was noted using the rule Relational data, and first data centers [3].

Looking back a little bit back to 2005, when the social media revolution such as Facebook and YouTube began sweeping the world, people realized the importance and value of data that users generate during their communication and the like. In the same year Hadoop was created and developed for big data storage and later analysis. Little by little, NoSQL started emerging and gaining popularity at that time. So Hadoop (and later, Spark) made it easier for users to store big data significantly, and they were open source tools as well, and the data is still flowing and increasing dramatically, not only by humans, but by others as well [3].

III. DISCUSSION

In this section, the frameworks and foundations of the term big data will be discussed, what are the security challenges and the tools used for that, which will explain the importance of this big data.

A. Big Data Basis

When talking about the concept of big data, which is often in the framework of analyzing big data instead of using the traditional method of data processing. We have been talking about speed, size, and diversity, and then value and honesty, which are characteristics that require complex methods and techniques that differ from traditional methods. One of the most important changes and differences is the dependence of that big data on new tools and techniques for the purpose of obtaining valuable information for institutions after processing it in the correct and modern technical ways, through the ability to link multiple data with each other. The term big data has replaced the old way of storing data in organizations, which resulted in a more comprehensive concept of data transfer and excellent results for companies and institutions business. This term has reached cloud computing and a wide spread of big data technologies, through the characteristics [5].

B. Big Data Characters

The main important characteristics that play a big role in big data are:

- Speed: While collecting large amounts of data and analyzing it quickly, the concept of speed and its importance is clarified through the enormous increase in data, such as images, videos, e-mails, social media, etc. Short and convenient time.

- Volume: It gives the concept to the amount of data that are generated every second is out of websites, mobiles, social media, credit cards, photos, videos, etc. The volumes of data streaming have become so large and expensive, in fact, it is very difficult to store and analyze data with the old traditional database technology. Currently, the last solution is to use distributed systems technology, in this new technology, pieces of data are stored in different locations and then collected and delivered together when needed using some software.

- Value: It indicates the value of the data that is being extracted. There is no point in having an amount of data unless it is converted into a value. There is a relationship between data and insights, So it is part of a project within the framework of big data, which helps to know the costs and benefits of collecting and analyzing data for a guarantee purpose the value of owning big data.

- Diversity: The term diversity of big data defined as the different types of data used. The data looks very different these days compared to the data from the past. You no longer have only structured data like name, phone number, address, financial data, etc., that fit well into your spreadsheet. However, the data today is disorganized. More than 80% of the world's data including photos, videos, social media, etc. Big data technology is innovative allowing structured and unstructured data to be collected, stored, and used simultaneously.

- Truthfulness: When talking about data accuracy and reliability, what is meant is data integrity and quality [6].

C. Big Data Tools

When talking about business for institutions and companies, we cannot ignore the term big data, which has become so important, which made those companies looking for specialists in the field of big data and their storage and analytical tools, which will be mentioned in the following:

- Apache Hadoop

It is not possible to talk about big data processing tools without talking about Hadoop, the popular primary data processing tool. Its use is suitable for open-source big data, and it can be used internally or even in the cloud depending on the hardware requirements. Hadoop has many advantages and benefits, including:

1. Hadoop Distributed File System (HDFS) has always worked at a wide frequency range.
2. B. Graphical model (MapReduce) for large data manipulation.
3. T. Yet Another Resource Negotiator for Hadoop Resource Management.
4. Enable the use of third-party modules by (Hadoop Libraries) to work with Hadoop.

- Apache Spark

It can be said that Apache Spark has become a replacement for Apache Hadoop. To handle Hadoop errors, Apache Spark was built which can process data in memory, and if compared to MapReduce it is much faster in disk processing. It is noted that Apache Spark Apache Cassandra, HDFS and OpenStack works in the cloud or wherever.

- Apache Storm

Processing data flow in time clarifies the concept of a storm and is supportive of many programming languages. Based on the architecture, it balances the workload between the different nodes, and works perfectly with Hadoop HDFS. Among the advantages of Apache Storm is its highly scalable horizontal scalability, and with the graph topology, it also works directly (DAG). When any malfunction occurs, it automatically restarts, works with output files in JSON format, as well as built-in fault tolerance directly.

- MongoDB

MongoDB is feature rich, as it is an open-source NoSQL database, and is also compatible with different programming languages, so it is a common platform for these languages. And when it comes to IT Svit which is the middleware for MongoDB to be used in a multitude of cloud computing monitoring and solutions. And if automatic MongoDB backups are to be used, the medium is Terraform. When talking about MongoDB's advantages, great flexibility in configuration, cloud deployment, and data segmentation across multiple nodes and data centers are among its most important advantages, as well as cost reduction as data is processed in one go through dynamic charts, as well as storing different types of data, an example of that number Correct, text, dates, matrices and more.

- Apache Cassandra

One of the most important components of Facebook's massive success is Apache Cassandra. Therefore, it is noted that data groups organized across many nodes are processed by Apache Cassandra in various parts of the world. Given that large institutions with heavy business are considered, this tool works well in those circumstances, relying on its structure without fail. If this tool is compared to NoSQL or DB, it surpasses them with its unique advantages, for example, the scalability of the font greatly, the continuous replication across the nodes, the use of a simple query language that helps to facilitate the operations, and, compact high availability, easy addition and removal of nodes, and High tolerance for errors [7].

D. Big Data Tools Security and Challenges Issues

- Data privacy

When talking about data privacy, which is one of the most important topics for users, and one of the most important and biggest concerns for companies and institutions that employ big data technologies and tools. As it is known that the big data system receives a large amount of personal information for the purpose of benefiting from it later, and this information is sensitive and expresses a private nature that the institutions have no right to use without informing the concerned parties. Although companies use this information for the purpose of profit benefits, but some data privacy techniques and tools help to encrypt that data and limit access to it, which reduces the opportunities for profitable benefits for these companies, as is the case between Apple and the social media giant Facebook, where Apple devices were prevented. Like iPhones, users can access data without their permission.

- Data management

The complete cycle of data gives it the concept of managing the data used in the big data system, an example of this, collecting data for the purpose of sharing it, and how to control the security of that data based on methods of collection and sharing, and the protocols followed for the purpose of securing the data.

- Integrity and security in interaction

The main purpose is to ensure that the data is original and has not been modified, by using various types of verification tools to ensure the integrity of the data. Therefore, it is important to know the extent of the possibility of discovering that the big data system may or may already be exposed [5].

IV. CONCLUSION

The main aim of this paper is to examine the role and concept of Big Data and its characters, tools, applications, and security challenges. Big Data is a great tool that makes things easy. Big data used in several applications such as banking, chemistry, agriculture, data mining, marketing, cloud computing, finance, stocks, BDA, health care, etc.

This industry has been growing every day. Big Data stays for the purpose of effectively influencing institutions and companies, the future of market share in terms of big data to increase manifold in next years.

REFERENCES

[1] Techopedia, "Big Data," Techopedia Inc, [Online]. Available: https://www.techopedia.com/definition/27745/big-data. [Accessed 25 December 2020].

[2] SAS, "Big Data," SAS, [Online]. Available: https://www.sas.com/en_sa/insights/big-data/what-is-big-data.html. [Accessed 25 December 2020].

[3] Oracle, "What Is Big Data?," Oracle, [Online]. Available: https://www.oracle.com/big-data/guide/what-is-big-data.html. [Accessed 25 December 2020].

[4] Barnard Marr, "What Is Big Data? A Super Simple Explanation For Everyone," Barnard Marr & Co, [Online]. Available: https://www.bernardmarr.com/default.asp?contentID=766. [Accessed 25 December 2020].

[5] M. A. S. E. F.-M. Julio Moreno, "Main Issues in Big Data Security," [Online]. Available: https://res.mdpi.com/futureinternet/futureinternet-08-00044/article_deploy/futureinternet-08-00044.pdf?. [Accessed 25 December 2020].

[6] J. Cano, " The V's of Big Data: Velocity, Volume, Value, Variety, and Veracity," [Online]. Available: https://www.xsnet.com/blog/bid/205405/the-v-s-of-big-data-velocity-volume-value-variety-and-veracity. [Accessed 20 December 2020].

[7] V. Fedak, "8 Open Source Big Data Tools to use in 2018," [Online]. Available: https://towardsdatascience.com/8-open-source-big-data-tools-to-use-in-2018-e35cab47ca1d. [Accessed 24 December 2020].