International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 9 Sep 2021, pp: 113-117 www.ijaem.net ISSN: 2395-5252

How efficient is big data in keeping private information safe when so many companies access it in real-time?

Prashast Pandey
The Scindia School, Gwalior, Madhya Pradesh, India

Submitted: 30-08-2021 Revised: 03-09-2021 Accepted: 05-09-2021

ABSTRACT: Big data has been playing a huge role in the information technology industry with about 50% of companies are collecting data and about 12% of them manipulating the data that they have, which may be used to analyse it computationally to reveal patterns, trends and especially human interactions and behaviour. Big data is often seen as a threat to the significance of other jobs, with the structured data used to feed AI and cultivate an ecosystem where humans are not entailed.The global **big data** market is estimated to **grow** to 274 billion U.S. dollars by 2027, more than five times its predicted market size in 2017.The proliferated use of big data has boomed the **data science** industry resulting in the increase in the number of **jobs** by about 28% through 2026. The big data is shaping the world we live in today, which could have endless possibilities.

Keywords- Big Data, Data Science.

I. INTRODUCTION

What is 'Big data'? Big data is often calculated by the erroneous units like Terabyte, Petabyte or Exabyte, but the big data can be of small size, like we cannot have an attachment of 100MB file in our email. Hence, that file can be referred as Big data with respect to email. We cannot imagine our world the way it is without the presence of Big data. Big data has been existent for a long time, way back in 1937. When Social Security Act became law in 1937, the government had to keep track of contribution from 26 million Americans and more than 3 million employers. IBM got the contract to develop punch card-reading machine for this massive bookkeeping project. This project laid the foundation of the big data and its other constituent industries that we know now. The emergence of Web 2.0 has made big data crucial and powerful, allowing augmentation of human interactions over the web. More and more companies are moving towards the big data to target their customers effectively. Big data has helped to change the way you see the web. Big data permits marketers to use subtle methods for presenting customer-specific content when and where it is most effective to improve online store recognition. A McKinsey survey found that "intensive users of customer analytics are 23 times more likely to clearly outperform their competitors in terms of new customer acquisition". Big data has made the usage of AI possible and both are interdependent in nature. The growing importance and the expansion of data science has made data breaches more vulnerable. Big data can be used by an evil entity easily leading to identity theft, blackmail, reputation or social damage.The protection of big data becomes an important task. When IoT is continuously rising at an incredible pace, the volume of big data rises exponentially, this equally aids the expansion of the cyber security domain. The data breach of integral datasets is one of the biggest digital threats anticipated in future. The qualitative real-time access of the big data makes make it vulnerable for interception.More data has been created in the past two years, than in the entire history of human existence and the emergence of big data has left people unaware about the responsibilities that come with them. The witnessed data breaches are less than 1% of the possible outcomes, with exabytes of data present, privacy of the entire world could be diluted. The objective of this research paper is to question the legitimacy and security of the real-time access of the big data harnessed by companies and pivot the attention towards necessity of protocols that need to be established to ensure the safety of the procured data.

1. **Mechanism of big data:** No computer can withstand the monstrosity of big data and that's the reason why we structure it, to add value to it and gain insight into the data. Big data can be broken down into five V's: volume, velocity, variety, veracity and value.

Volume:It refers to the amount of data that is encompassed in big data, usually generated by organisations or individuals.

Velocity:It refers to the pace of the data processing. The higher the velocity rate, the faster the data can be acquired and processed.

Variety: The various types of data are floating on the web and when stored, they can be classified into- unstructured, structured and semi-structed data.

Veracity:The accuracy or the trustworthiness of the retrieved data. **veracity** refers to the quality of data that is to be analysed.

Value: The meaningful insights or patterns gained from the dataset, which can be used for efficient business operations and other commercial benefits.

There is a farrago of big data tools available for analytics developers. Some of them are-

- **Apache storm**
- **Hadoop**
- **Casssandra**
- **Cloudera**

These tools are often used to extract information from a large number of data sets, with some of them providing real-time distributed system for processing data.

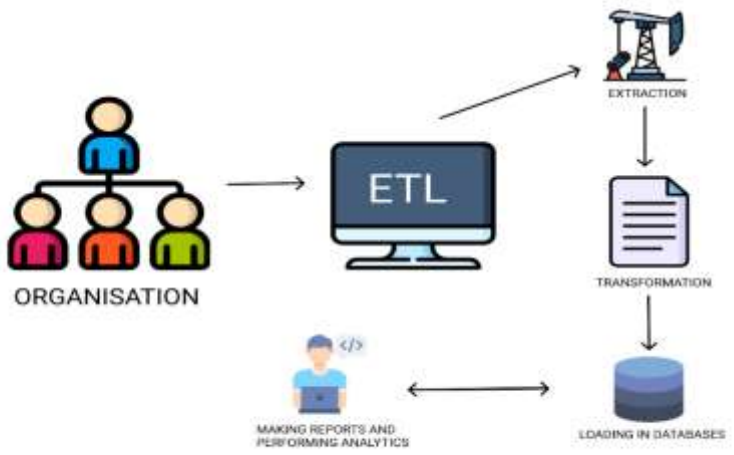
1.1 Why big data tools?

Can you think of processing exabytes of data from a traditional database? No, databases were not designed for it. The creation of big data tools like Cassandra enabled us to extract a large amount of data and process them. Data extraction is the process of obtaining data from a data base or SaaS platform and transferring it to a destination- such as a data warehouse, which can be used for online analytical processing(OLAP). **Hadoop distributed file system** is an integrated system of file distribution management that handles a huge amount of data sets running on commodity hardware. Let's understand the inner workings of both methods of processing big data for better understanding.

1.1.1 The traditional approach

The traditional method has been dominant for a long time. People were satisfied with this sluggish structure because of the meagre amount of data that has to be processed.

Collection of data: The data is collected from organisation such as hospitals, banks and corporate houses. The data is generally based on the recorded human interactions and related information.**Processing of data:** The data that is fetched from the organisations is fed into the Extraction, Transformation and loading system (ETL system). Wherein it transforms from unstructured data to structured data. Extraction process collects the information and converts it into unstructured data. Transformation process takes the extracted or the unstructured data and converts it into semi-structured data by formation of excel sheets, word documents or any other substantial form. The data of the documents are fed into databases for gaining an insight of the data.**Evaluvation of data:** An insight or value of the data is created in this process. The End-user gains the data that is loaded in the database and forms out reports and analytical resolution for the organistions with a specific objective.



Drawbacks:

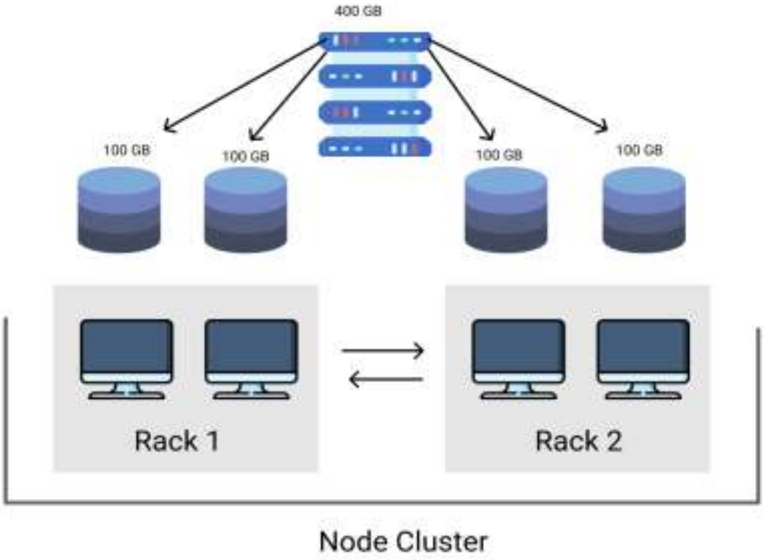
- It is very expensive. Hence, only wealthy organisations can have access to business analytics.
- Struggles with scalability. It cannot easily expand to a huge network.
- Consumes a lot of time to process.

As the data grows it becomes a challenging task to manage and process data by the traditional approach, as the big data surpasses the computation capacity of the machine. Henceforth, big data processing is quite different.

1.1.2 Big data processing

Big data processing is a series of techniques or procedures to access large-scale data to extract useful information for supporting and providing decisions. Big data is often unstructured and comes from different sources, makes it a poor fit for processing data in traditional databases. Therefore, big data tools allow us to process big data using a network of many computers which involve massive amounts of data and computation. Let's look into inner working of Hadoop (HDFS) for better understanding.

Hadoop does distributed processing for huge data sets across the cluster of commodity servers and works on multiple machines simultaneously. Hadoop is based on **master/slave concept**. A **daemon** is a program running on the background rather than interacting with the user. **NameNode** is the daemon running of the master machine. It is the **centerpiece** of an HDFS file system. **DataNode** daemon runs on the slave nodes. It stores data in the **HadoopFileSystem**. In functional file system data replicates across many **DataNodes**.**Yarn** divides the task on resource management and job scheduling/monitoring into separate daemons.**Resource Manager** – It runs on **YARN** master node for MapReduce. **Node Manager** – It runs on YARN slave node for MapReduce. The **rack** is a physical collection of nodes in our Hadoop cluster, the data is divided in blocks by mapReduce and then facilitates concurrent processing. These racks communicate between themselves as they store copy of the block data in the first rack, in case there is a loss of data due to technical complexities.**HDFS** stores the data while **MapReduce** process the data and Yarn divide the tasks.



II. CAN BIG DATA KEEP THE PRIVATE INFORMATION SAFE?

Big data is an incredible tool, which is probably a cardinal aspect of a flourishing organisation but can also lead to privacy risk if managed poorly. Management is the key with big data, with data getting bigger and bigger, management becomes relevant. If an organization stops using data because of the fear that it'll lead to security breaches, they'll be making a big mistake. Without big data, an organisation will have hard time understanding their customers and making data-driven decisions.

1.2 What's really going on with our personal information?

Privacy, a definition which is often misinterpreted globally, which leaves an array of speculated ideas floating around. The generalised idea of privacy is- **condition or state of being free from public attention to intrusion inter or interference with one's acts or decisions**.

2.1.1 **Data breaches:** Data breaches occur when information is accessed without authorization. In most cases, data breaches are the result of out-of-date software, weak passwords, and targeted malware attacks. Unfortunately, they can cost an organization a damaged reputation and a great deal of money.

2.1.2 **Data brokerage:** The sale of unprotected and incorrect data is considered data brokerage. Some companies gather and sell customer profiles, which contain false information that leads to flawed algorithms.

2.1.3 **Data discrimination:** Since data can consist of customer demographic information, organizations may develop algorithms that penalize individuals based on age, gender, or ethnicity.

In the modern digital landscape of today, where phenomenon such as the "filter bubble," and "personalized marketing" are on the rise, many individuals fear that they live with their **privacy**, particularly their online **privacy** in a state of constant decline.

III. BIG DATA AND PRIVACY MANAGEMENT

With the emerging Big data analytics, the concern about privacy comes into spotlight. Big data privacy involves properly managing big data to minimize risk and protect sensitive data. Big data comprises large and complex data sets, many traditional privacy processes cannot handle the scale and velocity required.

- Employ real-time monitoring

Since a privacy issue can happen at any moment, organizations should find a solution that monitors data in real-time

- Implement homomorphic encryption

Homomorphic encryption is a form of encryption that allows users to compute data without decrypting it first. This form of encryption should be implemented to store and process information in the cloud to prevent organizations from revealing private information to outside vendors.

- Avoid collecting too much data

Only the data that is absolutely necessary should be collected. An organization may not need the Social Security numbers of their customers; customer login usernames and passwords may only be necessary.

The privacy protection is determined by how efficiently can the company manage the big data that they are harnessing in their interests. The increase in the data requires the role of cyber security in the procuring and managing the data safely.

IV. CONCLUSION

Big data is like the ocean, no one can store the ocean but you can extract minerals from it and when you are responsible of a huge moana, you probably don't want to give it to the pirates and let them use your precious treasure. Now, that's the current big data scenario.

REFERENCES

1) Rahul Beakta "Big Data And Hadoop: A Review Paper" in Baddi RIEECE -2015, Volume 2, Spl. Issue 2(2015)BUEST.https://www.researchgate.net/publication/281403776_Big_Data_And_Hadoop_A_Review_Paper

2) Nada Elgendy and Ahmed Elragal "Big Data Analytics: A Literature Review Paper"German University in Cairo, pp. 214–227, (2014) . https://www.researchgate.net/publication/264555968_Big_Data_Analytics_A_Literature_Review_Paper

3) Elgendy, N.: Big Data Analytics in Support of the Decision Making Process. MSc Thesis, German University in Cairo, p. 164 (2013)https://cyberleninka.org/article/n/1428429.pdf

4) Lee, R., Luo, T., Huai, Y., Wang, F., He, Y., Zhang, X.: Ysmart: Yet Another SQL-to-MapReduce Translator. In: IEEE International Conference on Distributed Computing Systems (ICDCS), pp. 25–36 (2011)https://www.researchgate.net/publication/221460084_YSmart_Yet_another_SQL-to-MapReduce_translator

5) L. Guo, E. Tan, S. Chen, X. Zhang, and Y. E. Zhao, "Analyzing patternsof user content generation in online social networks," in KDD, (2009).https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.151.1621&rep=rep1&type=pdf

6) Serrat, O.: Social Network Analysis. Knowledge Network Solutions 28, 1–4 (2009).https://www.adb.org/sites/default/files/publication/27633/social-network-analysis.pdf

7) Zeng, D., Hsinchun, C., Lusch, R., Li, S.H.: Social Media Analytics and Intelligence. IEEE Intelligent Systems 25(6), 13–16 (2010)https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5678581

8) M. Wang, T. Madhyastha, N. Chang, S. Papadimitriou, and C. Faloutsos. Data mining meets performane evaluation: Fast algorithms for modeling bursty traffic. In Proc. of ICDE, (2002)http://www.cs.cmu.edu/~mzwang/research/pub/tr-101.pdf

9) Y. Sun, J. Zhang, Y. Xiong and G. Zhu. Data Security and Privacy in Cloud Computing.(2014)https://www.researchgate.net/publication/274230804_Data_Security_and_Privacy_inCloud_Computing

10) N. Leavitt, “Is cloud computing really ready for prime time?”Computer,vol.42,no.1,pp.15–25,(2009)https://www.researchgate.net/publication/274230804_Data_Security_and_Privacy_in_CloudComputing

11) Y. Duan, J. Edwards and Y. Diwedi. Artificial intelligence for decision making in the era of Big Data.https://uobrep.openrepository.com/bitstream/handle/10547/623124/Artificial%20Intelligence%20in%20the%20era%20of%20Big%20Data%20revised%20final%20%281%29.pdf?sequence=2&isAllowed=n