International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 05 Issue: 01 | Jan - 2021 ISSN: 2582-3930

# SECURITY & PRIVACY ISSUES IN BIG DATA

**Dr. Archana Sharma**
Associate Professor
IMS Noida

## ABSTRACT

Due to the rapid growth and spread of network services, mobile devices, and online users on the Internet, leading to a remarkable increase in the amount of data. Almost every industry is trying to cope with this huge data. Big data, the phenomenon has begun to gain importance. Big data due to various properties like volume, velocity, variety, variability & complexity put forward many challenges. However, it is not only very difficult to store big data and analyze them with traditional applications, but also it has challenging privacy and security problems. For this reason, this paper discusses big data, its ecosystem, addresses problem with security in big data and proposes protection strategies of big data security & privacy.

## KEYWORDS

Big data, Security, Privacy, cyber security,, authentication, encryption

## 1. INTRODUCTION

The term ‘Big Data’ refers to the collection of data in a huge amount that cannot be analyzed, processed & stored in a traditional method. In today’s era, there are multiple numbers of devices that generate data every second. Data has become an important asset for the company for its growth, companies analyze this data to make their business better for the customer & create more revenue.

With the introduction of big data, there are changes to how an organization stores data and have allowed them to develop a more thorough and in-depth understanding of their business, which implies great benefits. Big data is associated with 5 key concepts: *Volume, Velocity,* V*ariety, Veracity, Value.*

**Volume -** It defines the amount of data that is being collected by the organization from different sources, including IoT devices, social media, videos and more. In the past storing it would have been a problem but cheaper storage on a platform like data lakes and Hadoop have eased the burden.

**Velocity -** With the use of the Internet connecting to multiple devices the speed with which we get the data is huge. Datastream into the business at an unprecedented speed and must be handled in a timely manner.

**Variety -** In Big data, there is not only structured data but we deal with different types of data like unstructured, semi-structured data which consist of emails, documents, videos, images, etc.

**Veracity -** It refers to the quality of data. Because data comes from different sources, it's difficult to link, match, cleanse and transform data across systems.

**Value -** It refers to big data that must have some value. As if the data is applied then it gives some meaningful output.

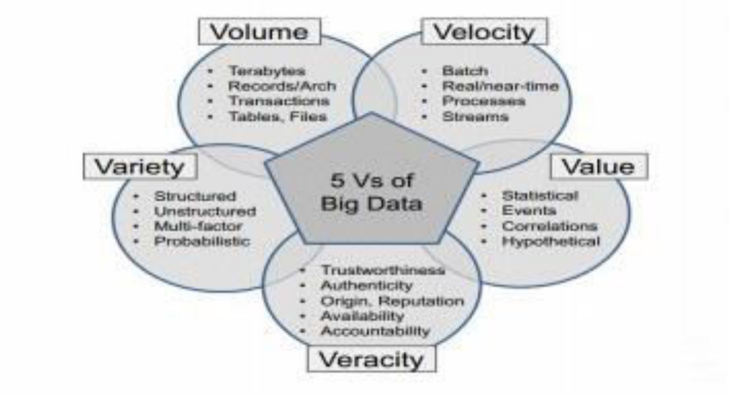

**Fig1:** *5 V's of Big data*

Above are the key concepts in big data where data is an important aspect. It is estimated that of all the data in recorded human history, 90% has been created in the last few years. In 2003, five exabytes of data were created by humans, and this amount of information is, at present, created within two days[1].

It is likely to be increased further at a rapid speed increasing the volume and the detail of the data collected by companies will not change in the near future. We are living in the era of Big Data where there is unstructured data which has to be processed and stored by the companies to do the analysis. It also signifies that the traditional method is no longer capable of analyzing it. As we have different forms of data then we have to handle the number of data increases from terabytes (TB) to petabytes (PB) and zettabytes (ZB). These huge amounts of data that complex have come to know as Big Data.

One of the most important parts of the big data is its technologies in order to extract valuable information from data and the ability to combine data from different sources and different formats. Big Data also helps change the way data is stored and develop them to get an in-depth understanding of their business, which implies great benefits.

## 2. BIG DATA ECOSYSTEM

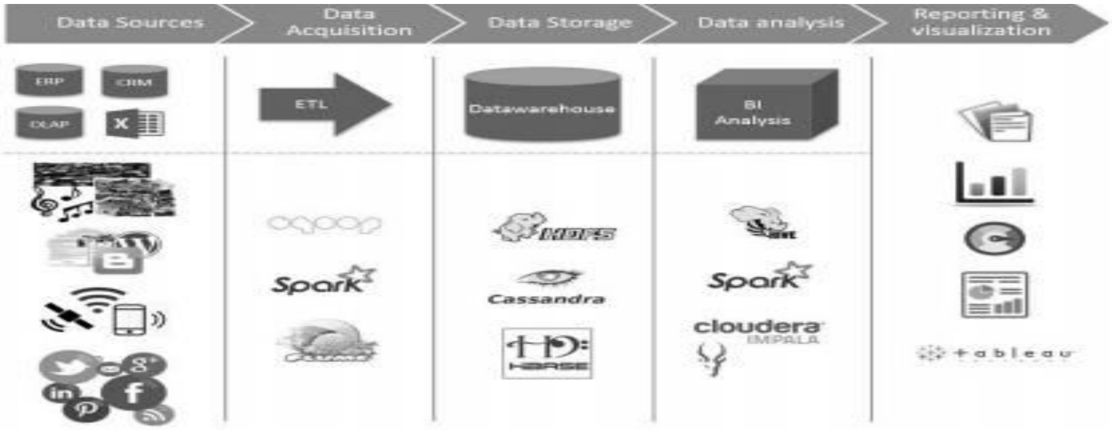

**Fig - 3** *Process in Big data*

Big data ecosystem contains many numbers of technologies that perform their own tasks from extracting the information until the visualization. Here first its data source from where we want to get the data like excel file, social media, ERP, etc then it goes through the ETL process (Extract Transform Load). The data is stored in a data warehouse like HDFS, HBase, etc which is suitable for NoSQL. The data stored in a data warehouse is further used for analysis for making the business decision. There are business intelligence tools like Spark IMPALA etc which display via a chart or through visualization tools like tableau etc. These processes are described in the above figure for better understanding.

## 3. SECURITY & PRIVACY

The use of Big data is increasing with the overflow of data in the huge amount and to select the related types of data from different sources which will help the organization for their business. Big data is used in many different fields like healthcare, internet, cloud computing, Social media, etc. Devices that generate a huge amount of data are connected to the internet and which leads to the risk of exposing the data to the hacker or third party. In big data, there are security and privacy issues that may expose companies' data which leads to damage to their business. Security & privacy are magnified by volume, velocity, variety and large scale cloud infrastructure. Big data is used with cloud computing in IaaS, PaaS, SaaS, etc providing the processing of data with storage until the visualization. In big data, there is a cybersecurity risk for the companies in which only 6% of companies are not exposed to cyberattacks while 62% of digital security threats[2].

### 3.1 Privacy

Privacy is a major concern in big data like in social media where it contains the huge amount of data of personal information of the user's which is compromised further leads to loss or leak of information which benefits the other from that data. The data shared in social media are provided to the organization which worries what the organization may do with that data. The termination of individual control over their own data is getting lost. Intelligence agencies like RAW, NSA, etc eavesdrop on the data of the organization and the individual in the name of national security. Big data projects like this show that there is a violation to the public interest.

The attacker stole the data and compromised the account of the user leading to the loss of money and data. Addressing these issues is not an easy task in data privacy. The integration of big data and cloud storage has caused privacy and security threats.

The following are the top security and privacy challenges given by CSA(Cloud security alliance) members studied publish research:-

- Secure computation in different programming frameworks.
- Security best practices for non-relational data stores.
- Secure data storage and transactional logs.
- End-point input validations/filtering.
- Real-time security.
- Cryptographically enforced access control & secure communication.
- Granular access control.
- Data provenance.

These above are the challenges that the CSA is mentioned which are of high priority. Data should be secure in the transaction as well as in storage[3].

## 3.2 Healthcare

Big data is being used in healthcare where security and privacy are a focal point as threats and vulnerability continue to grow. In healthcare, several factors provide the necessary force to harness the power of big data. Utilizing the power of big data in healthcare which provides doctors to make the decision in real-time to give medication to the patients. In recent times, upbringing in the newest technologies played the most significant roles in healthcare. For instance, realtime remote monitoring of vital signs through embedded sensors (attached to patients) allows health care providers to be alerted in case of any problem or difficult situation. Any loss or leak of the patient medical record may lead to patients dead or any serious problems. The use of big data technologies can enable us to get a deeper insight into the clinical and organization process but also facilitate the faster and safer throughput of the patient. It helps improve the patient engagement, safety, and quality of care of patients.

A dimensional insight study 56% of the hospital does not have proper medical practice data governance. Lack of proper data governance creates many problems in the organization, difficulty in finding a financial benchmark[4]. According to SOA(Society of Actuaries), big data is used in predictive analytics in analyzing that most of the cost in healthcare is afforded by them. Big data helps the clinic, hospitals, patients, etc to help find the patient to improve them in providing better medical facilities. A study published by Nature Review Drug Discovery found that 10% of medicine in development reached the patient. With the introduction of big data, the research and development of medicine are going in an effective way in short production time. Pharma companies save the cost of the drug in research and development because the process in which it is determined so that it can enter the hospital trial will be more accurate[5]. Big data helps in improved staff management which helps in who is working in shift work. The staff gets the improvement in managing the patient which will be less stressful so that they can manage the patient even better. They can also predict which kind of patient will be coming for treatment.

## 4. BIG DATA SECURITY & PRIVACY MEASURES

It is important to provide some measures in big data with security tools. Following are the technologies used to ensure security and privacy in big data [6].

- **Authentication -** It is an act of establishing or confirming claims made by or about the subject that are true and authentic. It serves vital functions within any organization: securing access to corporate networks, protecting the identities of users, and ensuring that the user is really who he is pretending to be. Authentication works the same where the user input the data and that data is evaluated.

When the user inserts their data then the system collects the data about the user. Even users don’t know that the system is collecting their data. When a request is sent to the system for authentication then it matches the previously collected data which is stored in the system. If the data is matched the access is provided by the system to the user but data is not matched then the alert is issued or the account access is denied. In big data, behavior analytics is done on the user who is accessing their account for determining the fraud or unauthenticated user. From which we can perform complex functions in managing the authentication of the system[7].

- **Encryption -** Data encryption is the best method of preventing unauthorized access. Encryption is the solution to protect the data from the theft from the data center to the cloud. To avoid any attack & breach encryption techniques are used to prevent packet sniffing and theft of storage devices. There are many encryption techniques that can be used to prevent theft such as RSA, Twofish, AES, and many more.

- **Access Control -** In this once the user is authorized then the data which is to be governed by the access control policy which ensures what the user can access in their account. It provides sophisticated control to ensure that the user can perform the activity for which the permission is given such as data access, job submission, etc.

- **Improve people’s awareness** - People’s awareness should be important because, in the end, they are one who is going to use the technology and share their personal information. The citizens should be data literate and aware of what data should be shared and what to not. The data awareness is directed at the people to make them understand the value of data and big data.

Don’t publish your data on the internet without understanding and other people’s information on the internet, so that they cannot be exploited by the criminals [8].

## CONCLUSION

In the era of big data it provides us the power to analyze, store and process the data easily in a huge amount which will help the organization in better decision making. In this paper, we have discussed big data, its uses and the domain where big data are used. If we have advantages then we have disadvantages also that are the security and privacy in big data where there is the loss of data and theft of the personal information from the account of the user. We have seen the problem in healthcare is using big data which compromises the data of patients which may lead to further losses[9]. So to overcome these problems we have given some measures in big data i.e. encryption, access control, improve people's awareness and authentication. By following these measures it is likely that the security and privacy theft of the personal information will be controlled. There are many more measures which can be used for security and privacy in further research.

In conclusion, big data is being used in many domains and help to solve their problems but everyone should be aware of the problem which may be caused by them.

## REFERENCES

1. Sagiroglu, S.; Sinanc, D. Big data: A review. In Proceedings of the 2013 International Conference on Collaboration Technologies and Systems (CTS), San Diego, CA, USA, 20–24 May 2013; pp. 42–47.

2. Alexei Balaganski & Dr. Sebastian Derwisch "Big data security analytics" october 2016.

3. Alvaro Cardenas Mora, Jesus Molina "Top Ten Big Data Security and Privacy Challenges" 2012 Cloud Security Alliance pp-5.

4. Nitin Kr. Agrawal "Big Data Security and Privacy Issues: A Review" Volume2 Issue4; September-October -2015;Page No. 12-15

5. Dongpo Zhang "Big Data Security and Privacy Protection, Advances in Computer Science Research", volume 77, 8th International Conference on Management and Computer Science (ICMCS 2018)

6. Anant Bardhan,Yu Chen, Adam Fuchs, Aditya Kapre in Big Data Working Group; Cloud Security Alliance (CSA). "Expanded Top Ten Big Data Security and Privacy". April 2013. (accessed on 9 December 2015) pp-5-7.

7. Workshop Report: Big Data Security And Privacy Sponsored by the National Science Foundation, The University of Texas at Dallas, September 16-17, 2014.

8. Wei Kaimin, Weng Jian, Ren Kui. A Survey of Big Data Security Protection Technology. Journal of Network and Information Security, 2016, 2(4).

9. Julio Moreno , Manuel A. Serrano and Eduardo Fernández-Medina "Main issues in big data security " 29 August 2016.