# A Review on Security of Big Data

Nikita Saxena and Dr. Hari Om Sharan,
*Department of Computer Science & Engineering,*
*Rama University, Uttar Pradesh, Kanpur, India*

***Abstract-*** With the fastest growing use of Information & Technology along with Internet, the amount of data is also increasing very rapidly day by day. So, the problem of managing that large amount of data has also been a crucial issue in the recent years. Big data analytics are the means which helps in generating new ways for many businesses and government industries to analyse and understand unstructured data. Now days, Big data is one of the most talked topic in IT industry. Big data helps in changing the way that the companies use to manage the large amount of data.

Big data comprises of data sets which are very large in size and also the data is complex in nature. Generally size of the data available in Petabyte and Exabyte. Artificial intelligence is often used as one of the technique to process this type of data. The potential benefit of Big Data resides in the fact that it has the ability to solve complex business problems and provide new business opportunities and insights. But without the consistency and reliability of the essential information, an organization loses its ability to make sound decisions. This paper defines the various concepts of big data and various security techniques performed on that large amount of data to maintain the security of that data.

***Keywords-*** Big data, Petabyte, Exabyte, Database, velocity, volume, variety, Analytics, Encryption, Cipher Text.

## I. INTRODUCTION

The term big data means extremely large data sets that may be analyzed computationally to reveal patterns, trends, and associations, especially relating to human behavior and interactions. **Big data** is a term used to refer to those data sets that are very large in size or are very complex in nature for traditional data-processing application software to adequately deal with. Data that consist of many cases (rows) offer greater statistical power, while data with higher complexity (more attributes or columns) may lead to a higher false discovery rate. A new term related with the same field and apparently more modern term is “Data Science”, which actually means the same thing, but includes the application areas to which “big-data” technology is applied.

The term Big Data came into existence around 2005 which refers to a wide range of large data sets almost impossible to manage and process using traditional data management tools – because not only of their large sizes, but also their complexity. There are many reasons for the increasing amount of data like use of internet, smartphone and social network. For managing that large amount of data traditional databases are not fruitful as they are not capable of managing, analyzing, and capturing that much amount of data. Due to the growing size of data, the security of data is also very important to maintain. For this many encryption algorithms have been used each having different encryption rates and processing time.

## II. CHARACTERISTICS OF BIG DATA

1) Volume- The volume defines the quantity of the data that is captured, generated or stored. This parameter defines that whether the data is considered to big data or not.
2) Variety- Variety determines what type of data is collected. It helps to analyse the nature of the big data to effectively use the insights of that data.
3) Velocity- Velocity defines at what speed the data has been generated to meet the coming demands of big data.
4) Value- Value of data depends upon whether the data is worth useful for the customers or not? That means whether a useful and required data can be found when needed or not?
5) Variability- This refers to the inconsistencies generated in the data time to time. These inconsistencies need to be managed properly in order to manage big data properly.
6) Visualization- Visualization means that whether the data makes any sense or not? Means whether the data helps in triggering any decision at a glance or not?
7) Veracity- Veracity of data refers to the accuracy of data.

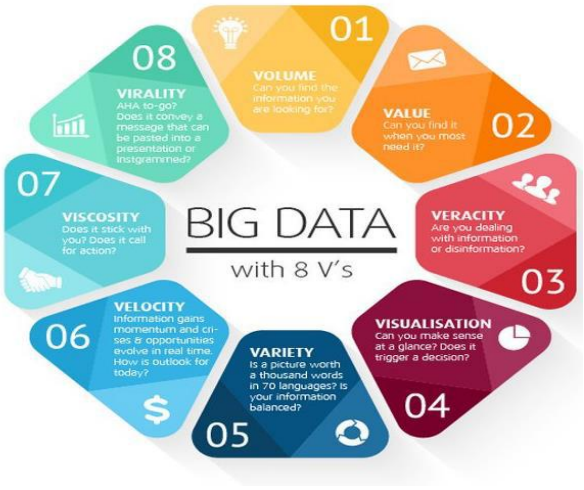


Figure 1- Characteristics of Big data

## III. STAGES IN BIG DATA

The stages involved in big data define the complete working of the big data. There are six stages in big data which includes the description of each and every phase of the big data process.

1. Data acquisition- It is the first and foremost important step of the big data. It involves the collection of large volume of data from various sources. With the growing use of internet, the rate of big data in also increasing exponentially. This large size data becomes more potent in nature when it's merged with other valuable data and superimposed. Due to the interconnectedness of devices over the World Wide Web, data is increasingly being collated and stored in the cloud.
2. Data Extraction- The second most challenging step in big data is extracting the required and useful data from the bulk of data collected in the first step. All the collected data is not useful for processing. So, it becomes very important to extract the required data.
3. Data Collation- While collecting the big data it, the data from the single source is not enough for processing. So it is important to get data from multiple sources.
4. Data Structuring- The structuring is important so queries can be made on the data. Data structuring is a way that employs methods of organizing the data in a particular schema. Various new platforms, such as NoSQL, can query even on unstructured data and are being increasingly used for Big Data Analysis.
5. Data Visualization- After converting the data into proper structure, it must be visualized into a format which is understandable and easy for interpretation of results.
6. Data Interpretation- The last step is finding proper outcomes and interpretation from the visualized data. This step is used for gaining valuable information from the data.

## IV. ANALYTICS OF BIG DATA

The analysis of big data is more important than big data. Big Data analytics is a science or technology to collect organize and analyse big data to discover patterns, knowledge and intelligence as well as other information within big data. Analytics of big data provides a convenient way for the organizations to use large amount of data and extract the required information from that data. Big data can be analysed for insights that lead to better decisions and strategic business moves. The most important aim of Big Data Analytics is to enable organizations to make better decisions with the help of better data. Due to the large Volume and Velocity of Big Data, data warehouses are unable to handle the complex processing demands posed by data sets that are being updated in real time environment and continually, such as the movements on social media websites. The newer technologies that are involved in Big Data Analytics involve Hadoop and related tools such as YARN, MapReduce, Spark, Hive and Pig as well as NoSQL databases.

## V. CHALLENGES IN BIG DATA

Due to the very large size of big data the challenges in maintain the security of big data has also been increased. Through recent disclosure, we now know that the NSA routinely collects and analyses massive amounts of personal data derived from heterogeneous data sources such as telecommunications, the Internet, and the user databases of large businesses, including Microsoft, Yahoo, Google, Facebook, YouTube, Skype, AOL, and Apple. Some of the most important security issues are as follows-

1. Privacy- Preserving Social Network Mining- Social networking is the major source of real life data. So, the mining of such type of data is very important concern.
2. Privacy- Preserving Big Data Analytics- Since analytics process huge volume of data hence the privacy of target data set is not preserved.
3. Security Aspects during Big Data Exchange- since the applications are exchanging bulk amount of data which becomes a clear source of security violation risks.
4. A major challenge for organizations is to choose and select the relevant and most important data from the data warehouse.
5. Security Issues of Big outsourced database- Since in cloud infrastructures, databases are very often outsourced which gives raise to very problematic security issues .
6. Privacy concerns when dealing with personal data, particularly for healthcare.

## VI. SECURITY OF BIG DATA

With the fastest growing demand and use of big data, its security concerns have also been increased. Securing these data has been a daunting requirement for decades. Various technologies are in use for protecting the security and privacy of data. Most widely used technologies are:

1. Authentication- Authentication is the process of establishing or confirming claims made by any individual or about the subject are true and authentic. It acts as a verification process protecting the identities of users, and ensuring that a user is who he claims to be.
2. Encryption- Encryption is the process that involves converting the plain text into cipher text. Encryption helps to prevent data from unauthorized access.
3. Data Masking: Masking replaces sensitive data elements with an unidentifiable value, but is not truly an encryption technique.
4. Access Control- Once authenticated, the users can enter an information system into the system. Thus it prevents an unauthorized access into the system.

## VII. CONCLUSION

The use of Big Data is increasing now days in a very large amount. So also the security concerns related to it are increasing. Big Data is a methodology to store very large quantities of data in an unstructured format and this is exactly what is needed to implement rich intelligent algorithms. The increasing use of Big Data analytics and AI in decision-making processes highlights the importance of examining their potential impact on individuals and society at large. The consequences of data processing are no longer restricted to the well-known privacy-related issues, but encompass prejudices against groups of individuals and a broader array of fundamental rights. Many new encryption algorithms and techniques have been emerging to overcome the security constraints of big data. But the problem of resource constraints in also a challenging factor in security of big data.

## REFERENCES

[1]https://www.researchgate.net/publication/275772328_Big_Data_Security_Issues_and_Challenges

[2] https://www.quora.com/How-are-Big-Data-and-Artificial-Intelligence-related

[3] https://www.marutitech.com/big-data-and-analytics/

[4]https://www.researchgate.net/publication/316679062_Big_Data_Analytics_and_Artificial_Intelligence

[5]https://link.springer.com/article/10.1007/s13218-017-0523-7

[6]https://www.sciencedirect.com/science/article/pii/S0267364918302012

[7]https://www.sciencedirect.com/science/article/pii/S1877050917317015

[8]https://journals.sagepub.com/doi/full/10.1177/2053951715609066

[9] https://www.renci.org/wp-content/uploads/2014/02/0313WhitePaper-iRODS.pdf

[10] https://www.sas.com/en_us/insights/big-data/what-is-big-data.html

[11] https://en.wikipedia.org/wiki/Big_data

[12]https://www.researchgate.net/publication/316644599_BIG_DATA_SECURITY_AND_PRIVACY_A_SHORT_REVIEW