A Review on Big Data: Privacy and Security Challenges

Parth Goel
Department of Computer Science & Engineering
Devang Patel Institute of Advance Technology and Research (DEPSTAR)
Charotar University of Science and Technology (CHARUSAT), CHARUSAT campus,
Changa 388421, India
er.parthgoel@gmail.com

Radhika Patel
Department of Information Technology
Devang Patel Institute of Advance Technology and Research (DEPSTAR)
Charotar University of Science and Technology (CHARUSAT), CHARUSAT campus,
Changa 388421, India
radhikapatel.it@charusat.ac.in

Dweepna Garg
Department of Computer Engineering
Devang Patel Institute of Advance Technology and Research (DEPSTAR)
Charotar University of Science and Technology (CHARUSAT), CHARUSAT campus,
Changa 388421, India
dweeps1989@gmail.com

Amit Ganatra
Department of Computer Engineering
Devang Patel Institute of Advance Technology and Research (DEPSTAR)
Charotar University of Science and Technology (CHARUSAT), CHARUSAT campus,
Changa 388421, India
amitganatra.ce@charusat.ac.in

2021 3rd International Conference on Signal Processing and Communication (ICPSC) | 978-1-6654-2864-4/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICSPC51351.2021.9451749

***Abstract*— In the age of technological innovations, the amount of data is increasing to a great extent. With this, an increasing trend is observed in the field of big data in industries as well as science. The scientific and industrial values of big data are growing high up to large magnitude. The technology of big data can be derived in many applications, but the main concern is the issue of security as well as privacy of data. This paper discusses the dimensions of big data and surveys the current research carried out on security as well as privacy of big data. The issues and the factors affecting the security are discussed. Furthermore, privacy-preserving approaches are also discussed and elaborated.**

Keywords*— Big data, security, privacy, privacy-preserving, risk analysis.

I. Introduction

Big data has begun as a new model or a prototype for numerous data applications. Industrial fields such as telecom, banking, healthcare sector, education sector, transportation, etc. [1, 2]. Big data is given attention along with data storage, data analysis, and data mining. Nevertheless, increased use of big data as a solution and data analysis mechanism doesn't promise security or/and privacy of the data. Besides the improvement of technology with the help of big data, privacy as well as security issues need to be taken into attention. Big data helps to preserve security issues with the help of security tools namely event management, network monitoring and security information [3]. Challenges being faced by big data in the field of security are the use of cryptographic algorithms, data provenance, security of stored data, access control, monitoring of real-time data etc. [4]. It is essential to analyze and identify privacy and security issues which will enhance the use of big data. The three principles of security are Confidentiality, Integrity and Availability. Security can be defined as the facility for monitoring person-specific access information, guard it from unauthorized revelation, alteration, harm or demolition of user's information [4]. Security can be obtained through controls based on operational and technical aspects. Privacy can be termed as a right of an individual to keep his/her personal information from being disclosed. Privacy can be obtained through policies as well as procedures [4]. Person's personal information which may lead to his identification may not be disclosed under ethical grounds. Security as well as Privacy of Big Data is essential in the applications of Bigdata such as healthcare, weather forecasting etc. This paper comprises of detailed information of research on issues of security as well as the privacy in big data. The focused data is trajectory data, i.e the data that represents the mobility of the objects. Attackers are most likely to attack the private data such as personal habits or personal data of mobile objects [5]. So, trajectory data mining has become an interesting topic for researchers to work on. The cryptographic algorithms, signature-based schemes, etc. are used to preserve the security of this trajectory data. It has been observed that to achieve big data privacy and security, infrastructure security, management of data as well as the integrity of data need to be taken into consideration. Moreover, the challenges being faced in achieving the issues of big data privacy as well as security is discussed in section II and III of the paper.

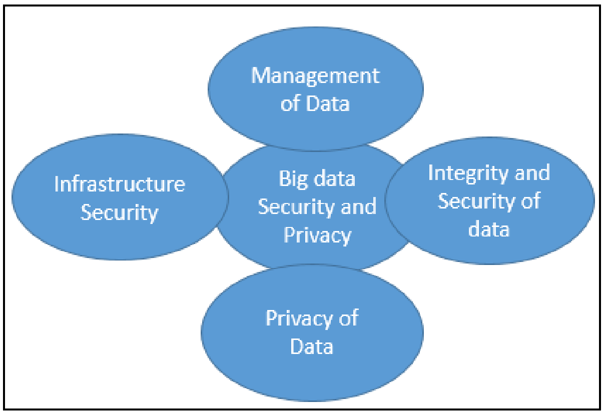

Fig. 1. Big Data Security and Privacy

II. Dimensions of Big Data

This section discusses the big data characteristics that are summarized as Volume, Variety, Velocity, Value, and Veracity [6]. These characteristics trigger the privacy and security of big data as discussed below.

1. “Variety” describes the diversity in the data format and the multiple sources through which data is been collected. The data can be structured, unstructured, or semi-structured whereas the data file types can be text, figures, or videos. Only the large-scale infrastructure of the cloud is capable to manage this bigdata. The security and privacy of cloud infrastructure are a challenge for the diversified storage of data.

2. “Volume” defines the size of the data. The amount of data that is produced every second by organizations, entities as well as sensors. The volume of the big data renders the risk of data leakage. Above and beyond, prevailing infrastructure approaches, like regular tracking, auditing, monitoring, and security scanning technology are insufficient as it becomes more complex and expensive to implement these methods on the huge amount of data [7].

3. “Velocity” leads to continuity and the high frequency of data. This feature makes privacy and security a little difficult to achieve. Vastly increasing and reiterating data needs non-relational databases which fail to set itself in any framework and thus becomes difficult to provide security as well as privacy aspect to big data.

4. “Veracity” means trustworthiness, bias, noise, applicability, and other qualities and properties of big data [8]. The veracity characteristics/feature involves the entire chain of big data starting from original data authenticity, the integrity of mined data, and the reliability of the data that is already published.

5. “Value” points to the output obtained from the huge sets of data. The integrated data obtained as an output generally attract hackers. This increases the risk of attacks on the database. Corporations, individuals, and organizations gain high benefits from predictive analysis of big data but at the same time, it is being used by cyber attackers. So, the tradeoff between preserving privacy and the benefits gained from data utilization is to be considered seriously [8].

Thus, the five Vs of big data trigger the security as well as privacy of big data.

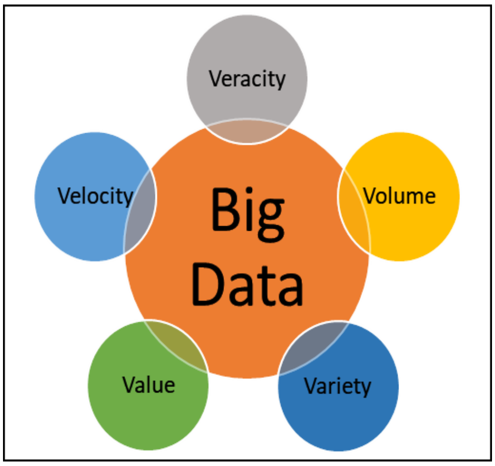

Fig. 2. Dimension of Big Data

Few more issues harm the data security in the big data ecosystem which is elaborated below.

1. Nowadays, the integration of big data and data storage on the cloud has become a challenge to

security as well as privacy threats of big data [12]. This is due to inefficient security applications which fail to provide security to dynamic and big volume data.

2. Protecting transaction logs, data and other sensitive information are also a challenge as the data may have varying levels.
3. Validating and filtering end-point input devices is also a challenge as the end devices are the main features through which the big data ecosystem is maintained.
4. Granular auditing and access control which is provided by databases like NoSQL or Hadoop Distributed File System necessitate a very robust process of authentication and compulsory access control [12].
5. Security of distributed framework like Hadoop is essential and is a challenge because the functions like MapReduce is used for mapping of the data.

III. Concerns in Big Data

Existing security, as well as privacy solutions for big data, are discussed in this section. The researchers have proposed some theoretical and functioning classifications of privacy and security to get familiar with vulnerabilities of big data. Security, as well as privacy, are required on four different layers of big data system viz. at storage layer for safe storing and controlling apparatus of monitoring. Second at management layer which includes Hadoop Distributed file system (HDFS), encryption of data, etc. [12]. The third is the interface layer which includes an application programming interface (API) that is public, identity authentication, and access management [12]. Fourth is at access layer which includes the cybersecurity of the users. The survey focuses on analyzing vulnerabilities as well as risks that are increasing day by day in the era of big data. Methods to enhance the security and privacy aspects are surveyed and discussed. Moreover, it has been surveyed that security needs to be given on data sources, stored data, and output data. Providing security to the above-mentioned categories of data will in turn help to provide security to big data.

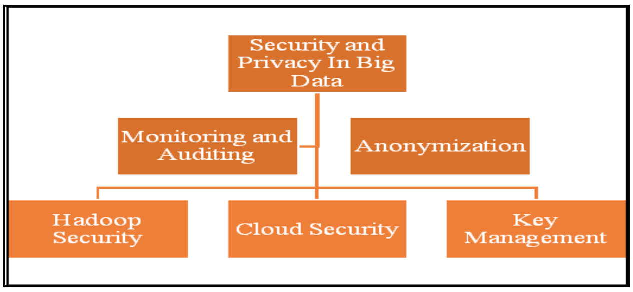

Fig. 3 Security and Privacy Categorization

Various schemes of data encryption, permissions to access, security of transport layer is easy to break and firewalls are easy to re-identify and alter. Due to these reasons, advanced technologies and techniques to provide security & privacy to big data are developed. The details as shown in Fig. 3 are described in the next section of the paper.

Security in Hadoop

Hadoop is regarded as one of the distributed process frameworks which are not designed for security. Hadoop is a popular platform for big data analytics which requires security. Two security techniques were proposed to avoid hackers from hacking the data from the cloud. A mechanism of trust is implemented between the name node and the user that is a component of HDFS. In the mechanism, the user needs to authenticate itself to the name node. The hash function is produced together, by the user and the name node. Both the hash functions are compared. If it matches then the user is allowed to access the big data system. Hashing technique named SHA-256 is used in this technique for authentication. HDFS security has a significant value. So, three methods are developed for ensuring security. Kerberos mechanism dependent on Service Ticket is the first method used for security. The second method is related to monitoring sensor data and sensitive information by an algorithm named the Bull Eye algorithm [4]. The main advantage of this algorithm is to manage the relation between original data and replicated data. The Master slave method is used as a third method to ensure the security of HDFS. The master node and the slave node are two nodes working as a name node in HDFS. If there is a problem in the master node, the slave node can react and answer with the permission of the administrator using Name Node Security Enhancement (NNSE) permission.

Security on Cloud Platform

Broad network access, on-demand services and resource pooling are some of the reasons of widely spread of cloud computing usage. The hosts in the cloud architecture prevails the threat of attacks. Therefore, the service provider of the cloud architecture needs to take the precaution for the same. Security techniques such as authentication, compression, encryption as well as decryption etc. are used to secure big data on the cloud platform. Cryptographic Virtual Mapping is used in a security solution mechanism on the cloud platform for the creation of data path. In the proposed mechanism, security is provided to the information which seems to be critical, sensible and useful. The encryption is done only on the storage path which shows the direction of the critical, sensible and useful data. To achieve the factor of avaibility, there are numerous copies of each data part as well as their accessing index [12]. Thus, if part of the data is lost, then information availability is maintained successfully.

Monitoring as well as Auditing

Monitoring and auditing point towards gathering and investing about the network events to detect intrusion [12]. Security monitoring architectures like DNS traffic, HTTP traffic, and others are built because it seems very difficult to detect intrusion on the whole traffic network. The correlation scheme of data is utilized to accumulate and process data in the disseminated sources. A matrix is prepared to know whether the node, flow, or packet is malicious or not. The alert message arrives in the detection system if any of these is found or the process is terminated through the prevention system [12]. Data Availability, Integrity, Consistency, Aggregation, and Confidentiality has emerged as a gap in big data security along with big data characteristics. Therefore, security solutions need to be applied for filling this gap.

Key Management

Sharing and Generating keys between the server and the user also lead to a security issue in big data security. There is a need for a group key transfer protocol to share a key between multiple groups. Therefore, to avoid the online threat of attackers, a novel protocol is established with the help of the Diffie-Hellman key agreement as well as a linear secret key mechanism. To securely share the data, the utilization of the Outsourcing Conditional Proxy Re-Encryption (CPRE) is done in a group of complex networks. Moreover, the security of unstructured data like email, text, XML is difficult in big data systems [4]. Filtering, clustering as well as classification centered on data sensitivity level in the phase of data analytics can be used for securing unstructured data. In the second level, the data node of the database is organized and appropriate service (identification, integrity confidentiality, non-repudiation, and authentication) is selected by a scheduling algorithm from the security suite to provide security.

Anonymization

Various classical methods are used to fill anonymity over big data but none of the methods are capable to provide security due to the volume characteristics of big data. The hybrid scheme to handle anonymization is proposed that combines two classical approaches namely Bottom-Up and Top-Down for Sub-tree Anonymization to increase scalability and privacy. Another scalable method is introduced to solve the problem of scalability wherein the dataset is split using the t-ancestor clustering method and then data is recorded using a proximity-aware agglomerative algorithm [12]. A differential privacy approach is presented to preserve privacy in big data. The model is designed in such a way that ensures an equal possibility of data to get released amongst all input data. Two mechanisms known as the Laplace mechanism and the Exponential mechanism are designed for providing differential privacy in big data [12, 13]. For actual and up-to-date outcomes, the Laplace mechanism is utilized where the noise generation takes place built on Laplace distribution. Whenever outcomes are unreal, the Exponential mechanism allocates exponentially better possibility to an output with a greater score. With this, it is more possible to be designated [12]. The differential privacy is verified to consume healthier isolated usefulness for trajectory data; nevertheless, the research or study should consider adjusting the gigantic bulk of real-time data. Moreover, it has been surveyed that security needs to be given on data sources, stored data, and output data [12]. Providing security to the above-mentioned categories of data will in turn help to provide security to big data. Machine learning and statistical analysis which is collectively known as data science techniques are some of the techniques used to adapt to the updations required in the environment of security. With the help of this technique, analysis is done based on the type of data in the ecosystem and various machine learning algorithms are used to detect any changes in the data without the help of any human intervention [13]. There are various other security tools used by the companies in the market presently to provide security to the big data ecosystem. IBM risk manager tool is of them which gives facilities to manage network device configuration for compliance reporting and risk management [14].

Figure 4 depicts the issues of security, attacks as well as their possible resolutions concerning Infrastructure security, data privacy, data management, integrity, and reactive security [14].

IV. Conclusion

Big data has been proved to be the most favorable and prevalent technology for forecasting the trends of the future. In this situation, privacy and security require to be considered the topmost priority for sensitive applications in big data. Thispaper analyses the big data characteristics effects on the security as well as privacy on the infrastructure of big data systems, cloud security, and data management. The next section of the paper surveyed the research trends on big data security as well as privacy. Various mechanisms concerning security and privacy were discussed and compared for solving the issues of security as well as privacy in big data. The focus was on the latest security as well as privacy mechanism which proves efficient to offer security as well as privacy in complex networks of big data

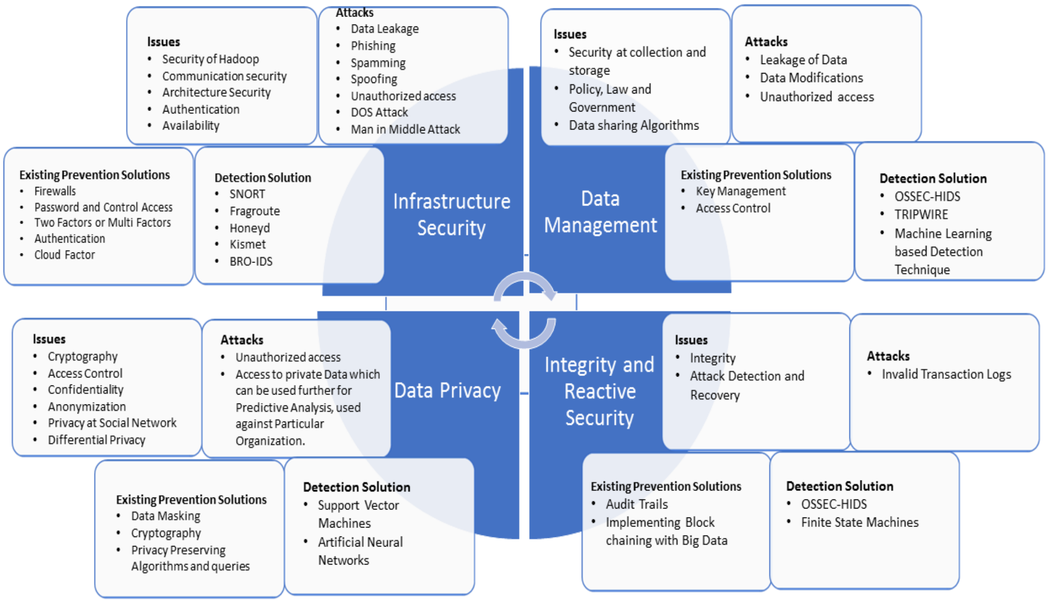

Fig. 4. : Security and Privacy issues and attacks along with soutions in Big data

REFERENCES

[1] X. Cheng, L. Xu, T. Zhang, Y. Jia, M. Yuan, and K. Chao, "A novelbig data based telecom operation architecture," in 1st International Conference on Signal and Information Processing, Networking and Computers, 2016, pp. 385–396.

[2] L. Xu, Y. Luan, X. Cheng, X. Cao, K. Chao, J. Gao, Y. Jia, and S. Wang,"Wcdma data based lte site selection scheme in lte deployment," in1st International Conference on Signal and Information Processing,Networking and Computers, 2016, pp. 249–260.

[3] A. A. Cardenas, P. K. Manadhata, and S. P. Rajan, "Big data analytics for security,"IEEE Security & Privacy, vol. 11, no. 6, pp. 74–76, 2013.

[4] D. S. Terzi, R. Terzi, and S. Sagiroglu, "A survey on security and privacy issues in big data," in2015 10th International Conference for InternetTechnology and Secured Transactions (ICITST). IEEE, 2015, pp. 202–207.

[5] Y. Deng, L. Wang, S. A. R. Zaidi, J. Yuan, and M. Elkashlan, "Artificial-noise aided secure transmission in large scale spectrum sharing net-works,"IEEE Transactions on Communications, vol. 64, no. 5, pp.2116–2129, 2016.

[6] H. Ye, X. Cheng, M. Yuan, L. Xu, J. Gao, and C. Cheng, "A survey ofsecurity and privacy in big data," in 2016 16th international symposiumon communications and information technologies (iscit).IEEE, 2016,pp. 268–272.

[7] B. Matturdi, X. Zhou, S. Li, and F. Lin, "Big data security and privacy:A review,"China Communications, vol. 11, no. 14, pp. 135–145, 2014.

[8] J. Qiu, S. Jha, A. Luckow, and G. C. Fox, "Towards hpc-abds: An initialhigh-performance big data stack,"Building Robust Big Data Ecosystem ISO/IEC JTC, vol. 1, pp. 18–21, 2014.

[9] E. Zeng, S. Mare, and F. Roesner, "End user security and privacy concerns with smart homes," in thirteenth symposium on usable privacy and security ({SOUPS}2017), 2017, pp. 65–80.

[10] C. S. Alliance, "Expanded top ten big data security and privacy chal-lenges," 2013.

[11] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise tosensitivity in private data analysis," in Theory of cryptography conference.Springer, 2006, pp. 265–284.

[12] F. McSherry and K. Talwar, "Mechanism design via differential privacy,"in 48th Annual IEEE Symposium on Foundations of Computer Science(FOCS'07). IEEE, 2007, pp. 94–103.

[13] C. Thota, G. Manogaran, D. Lopez, and R. Sundarasekar, "Architecture for big data storage in different cloud deployment models," inResearch Anthology on Architectures, Frameworks, and Integration Strategies for Distributed and Cloud Computing. IGI Global, 2021, pp. 178–208.

[14] R. Bhatia and M. Sood, "Security of big data: A review," in 2018 Fifth International Conference on Parallel, Distributed and Grid Computing(PDGC). IEEE, 2018, pp. 182–186.