Academic Guru Publication House

Advances in Cloud Computing Security
Techniques and Applications
Volume 1
Year: 2021

AG PH Books

A Review on Big Data: Privacy and Security Challenges

Dr. Rohit Kumar[1*]

[1]*Assistant Professor, IT, Institute of Management Studies Noida*

Abstract

A significant rise in data is being generated due to factors such as the fast expansion and dissemination of the network services, the mobile devices, but also internet users. Almost every business is attempting to deal with the massive amounts of data that are being generated. The concept of big data has starting to gain traction. Traditional apps can't handle huge data because it's hard to store as well as analyse, so it poses serious privacy as well as security concerns. As a result, this article explores the scope of big data and reviews the most recent studies on security as well as privacy issues surrounding it. We'll talk about the problems and the things that impact security. Privacy-preserving methods are also examined and expanded on in this paper.

Keywords: big data; Hadoop security; cloud security; monitoring; auditing; key management; anonymization.

1. INTRODUCTION

The term "big data" is a relatively new one in the field of information technology, and it refers to vast amounts of data which are now possible for collecting, storing, including managing, as well as analysing [1].

It is important to note that in [2], the term "Big Data" refers to data with a high volume, speed, and variety that must be processed in new ways to improve decision making as well as the process optimization. Data sets can be referred to as "Big Data" if their capture, analysis, storage, filtering and visualisation is beyond the capabilities of current or even such traditional technologies. As it's discussed

* ISBN No. 978-81-955340-6-7

through The Economist in [4], "Managed well, the data can be used to unlock new sources if economic value, provide fresh insights into science and hold governments to accounts".

With the use of security solutions such as network monitoring, event management, as well as security information [5], the big data helps in preserving security concerns. The usage of cryptographic techniques, data authenticity, security of the stored data, access management, and monitoring of the real-time data are just a few of challenges that big data faces when it comes to security [6]. Big data can only be used effectively if privacy as well as security concerns are addressed. Privacy, integrity, as well as availability are such three of the most important security concepts. When it comes to protecting user information, security may be described as the ability to monitor and protect user-specific access information against unauthorised disclosure, change or destruction [6]. Controls based on the operational and technological elements may provide security. In other words, the right to the privacy is the ability of the individual to limit the disclosure of personal information about themselves. Policies and procedures may be used to ensure privacy.

2. FEATURES OF BIG DATA

Volumes, velocities, varieties, truthfulness, and value are the five vs of Big Data characterization.

1) **Volume (data in rest):**There are two main characteristics of big data: volume (information in the rest) and scalability.
2) **Variety (data in many forms):**There are three sorts of data in the world: internal, an external, and a combination of the two
    - Clearly defined (information ffrom the relational databases)
    - Structured in a semi-structured manner (website logs, the social media feeds, the sensor data, email and so on)
    - A lack of structure (videos, images, audios and so on.)
3) **Velocity (data in motion):** "The production rate of data is notably high. The increase in data means that data should be analyzed more swiftly" [7]. Indeed it is not just the velocity of incoming data that is the issue it is possible to stream fast moving data into bulk storage for later batch processing.
4) **Veracity:**An issue of veracity is degree to which data are uncertain or inaccurate.
5) **Value:**It's really value of knowledge that can be gleaned from it.

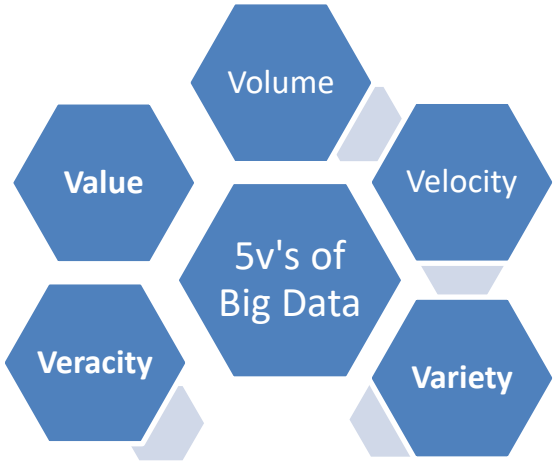

Figure 1: V's of big data

3. CHALLENGES OF BIG DATA

New value intuitions are expected to be gleaned from accessible data sources when Big Data is used. These include issues like the heterogeneity, the data life cycle management as well as processing, the data scalability as well as the security including privacy as well as data visualisation:

A. Heterogeneity

There are essentially endless dispersed data sources, which contributes to Big Data's variety [8]. Images, videos, audios and sensors are all examples of data diversity [9]. Because of such difficulties in converting structured as well as semi-structured information into homogeneous data, handling heterogeneous data types as well as sources is a critical future problem [10]. Metadata is also part of Big Data. The smartphone's metadata can tell you where a picture was taken, when it was shot, and even what kind of camera it used. The problem of dealing with a variety of metadata types is likewise a difficult one.[11].

B. Data Processing

Pre-processing may be necessary before the actual examination of data obtained from multiple sources. Data reutilization, data reorganisation, and even data exhaust [12] may be used to process redundant data for future study. It is possible that inaccurate data was created during the data mining in an attempt to improve mining results.

C. Data Life Cycle Management

There is no end to the data life cycle in the Big Data. Dataset values might well be interpreted in a variety of ways by various users. When a patient's recovery is complete, the patient's health record is no longer useful from the patient's viewpoint, but it might be useful from the doctor's or even researcher's perspective [9, 12]. Because of this, it is necessary to revisit how data lifecycles are judged and defined.

D. Security and Privacy

Conventional protection methods won't work for Big Data due of its unique properties. Because of this, the widespread usage of the Big Data in everyday operations raises security concerns. Outsourcing sensitive information also raises concerns about data security. Data capture or even data storage of the personal sensitive data must be done in a manner that protects the privacy of such data. [13], [8-9], [12].

E. Scalability

Volume has a direct bearing on scalability. Storage scalability refers to the system's capacity to accommodate growing data volumes in an appropriate way [8]. As the phrase "Big Data" implies, it necessitates an ever-increasing number of data to be processed.

4. BIG DATA SECURITY AND PRIVACY APPROACHES

Security as well as privacy are not adequately protected when working with large amounts of data. When it comes to encrypting data or accessing it, access restrictions, firewalls, the transport layer security, and even the anonymized data may all be compromised [14]. This is why improved methodologies and technologies are being created to safeguard, monitor as well as audit big data activities from an infrastructures, applications as well as data perspective. Based on previous research on these topics, this article has grouped security and privacy concerns for big data into the following five titles: cloud security, Hadoop security; monitoring as well as auditing; key management as well as anonymization (Figure. 2). To summarise the research, we created Table I, which lists the studies' aims, methods, and results.

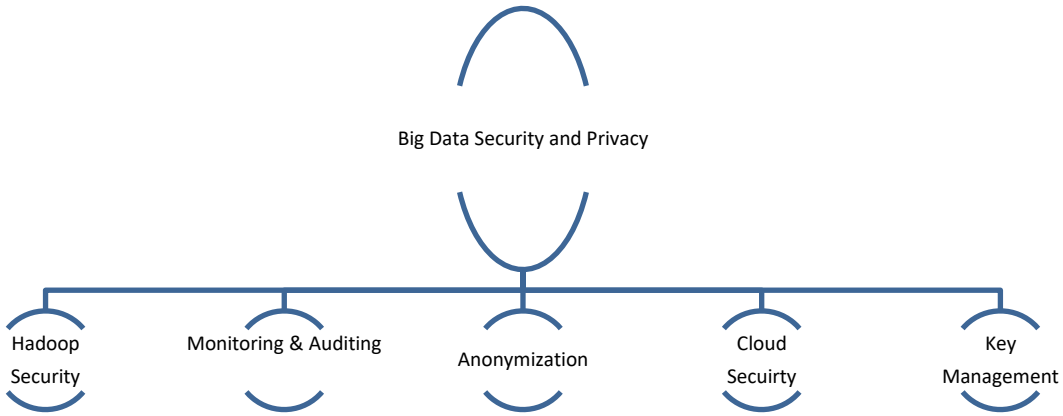

Figure 2: Big data security and privacy categorization

a) Hadoop Security

Hadoop wasn't really designed with security in mind when it was first built as the distributed process platform. It's designed to work in safe surroundings. The development of security measures for Hadoop has accelerated in response to its growing popularity. In addition, academics are starting to pay attention to it. To prevent hackers from accessing all of such data stored within Hadoop system, two solutions were presented [15]. The name node, that is part of HDFS and controls data nodes, has created the trust mechanism between the user and the node. For accessing a name node, the user should authenticate using this technique. Firstly, the user provides the hash function, as well as then the name node generates a hash function of its own including compares them. Access to the system is granted if the comparison results are valid. One of the hashing algorithms is employed in this phase to authenticate the data. In order to prevent hackers from accessing the whole database, random encryption algorithms including , AES, RSA, Rijndael and RC6 have been utilised. This technique uses MapReduce to do the encryption or decryption operation. Ultimately, twitter stream is used to test such two methods and show how to manage security concerns.

b) Monitoring and Auditing

Gathering and studying network events in order to detect breaches is the goal of network security monitoring. It is the systematic and quantifiable security strategy to apply several approaches for security audits. Active security relies heavily on both of these factors.. On the overall, network traffic detection as well as prevention techniques are quite tough. There is a solution to this issue by analysing the DNS traffic, the IP flow records, the HTTP traffic, and the honeypot data [16]. Data correlation techniques are used to store and analyse data from several sources in a dispersed manner. In order to determine if the domain name, packet, or even flow is malicious, 3- probability metrics have indeed been computed. An alarm is triggered in the detection system or the procedure is terminated mostly by the prevention system based on just such score. Using open-source big data platforms, Shark and Spark outperform

Hadoop, Hive, as well as Pig in terms of speed and consistency when it comes to analysing electronic payment data from a corporation.

c) Anonymization

Data mining for analytics raises serious issues about privacy. When it comes to protecting personal information (PII), it's becoming tougher. Agreement between the firm and person must be based on rules to eliminate privacy problems. De-identification and anonymization of personal information are critical steps in the data transmission process [17]. However, the company's algorithms and artificial intelligence analyses may reveal the individuals identify. This analysis's conclusions have the potential to lead to unethical situations.

d) Cloud Security

Due to factors like on-demand service,the pooling of resources, and flexibility, cloud computing has become a suitable setting for big data [6]. It is now common practise to employ the cloud computing. The cloud, on the other hand, is vulnerable to both old and new dangers. In today's world, the cloud data storage is indeed a major issue. As a result, service provider should take certain measures to protect its customers. As a result, a safe method for managing and sharing large amounts of data on the cloud has indeed been developed [18]. Authentication, Security measures such as decryption, as well as compression are included in order to protect large amounts of data. For the authorised user, email as well as password authentication was utilised. In order to ensure the safety of the data, it has indeed been compressed as well as encrypted. In the event of a natural catastrophe, the company employs 3 backup servers. In such servers, encrypted data has indeed been saved. The secret key was used to decode encrypted data in event of a server failure.

e) Key Management

Another important data security concern is the generation and distribution of keys between servers as well as users. Fast as well as dynamic authentication procedures, on the other hand, may be proposed for large data centres. The PairHand protocol was developed for authentication in the mobile or stationary data centres in [19] using a tiered approach to the quantum cryptography, data-reading, the front-end, , quantum-key-processing-management, as well as application layers all fall under this umbrella term. Both key search operations as well as passive assaults have been decreased by this approach.

The big data services are made up of a number of different groups, each of which requires a safe way to exchange group keys. As a result, a new protocol-based o the Diffie-Hellman key agreement as well as a linear secret sharing method has indeed been proposed instead of the current protocols [20]. In order to protect the system, protocol ensures freshness of key, authentication of key, and secrecy of key.

CONCLUSION

New signs of scientific revolt are on the horizon as we move into an age of big data that represents next frontier for the revolution, competition, as well as efficiency. There are many issues with Big Data, and we've covered them all in such article, from heterogeneity to management of the data life cycle to data processing to privacy as well as security. Security including privacy problems like data privacy as well as key management have also been thoroughly examined. We have examined them in-depth. In order to address the above-mentioned security including privacy concerns, a number of commendable efforts have been made. Future discussions on big data security, privacy, as well as safety will need new methodologies as well as technologies for the human-computer interactions or even the improvement of current ones for the more accurate outcomes. Research in these areas will, however, need careful attention and effort on the part of academics. We believe that our extensive poll will assist to build better security as well as privacy solutions for Big Data, which is in its infancy.

REFRENCES

1. Johns Hopkins, “Big data custodianship in global society”, SAIS Review of international Affairs, Volume 34, Number 1, Winter-Spring 2014, pp. 109-116 (Article).
2. C. Philip Chen and C. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data", Information Sciences, vol. 275, 2014, pp. 314-347.
3. D. Terzi, R. Terzi and S. Sagiroglu, "A survey on security and privacy issues in big data", 2015 10th International Conference for Internet Technology and Secured Transactions (ICITST), 2015, pp. 202-207.
4. E. Bertino, "Big Data - Security and Privacy", 2015 IEEE International Congress on Big Data, 2015, pp. 757-761.
5. 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.
6. 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.
7. Xinhua Dong, Ruixuan Li_, Heng He, Wanwan Zhou, Zhengyuan Xue, and Hao Wu, “Secure Sensitive Data Sharing on a Big Data Platform”, Tsinghua Science and Technology, ISSN l l1007-0214l l08/11l lpp72-80, Volume 20, Number 1, February 2015

8. I. Hashem, I. Yaqoob, N. Anuar, S. Mokhtar, A. Gani and S. Ullah Khan, "The rise of "big data" on cloud computing: Review and open research issues", Information Systems, vol. 47, 2015, pp. 98-115.
9. Weichang Kong, Qidi Wu, Li Li and Fei Qiao, "Intelligent Data Analysis and its challenges in big data environment", 2014 IEEE International Conference on System Science and Engineering (ICSSE), 2014, pp. 108-113.
10. [Online] Challenges and Opportunities with Big Data", Purdue Univesity, 2016. [https://www.purdue.edu/discoverypark/cyber/assets/pdfs/BigDataWhite Paper.pdf. [Accessed: 12- Jan- 2017].
11. Z. Azmi, "Opportunities and Security Challenges of Big Data", Current and Emerging Trends in Cyber Operations, 2015, pp. 181-197.
12. M. Chen, S. Mao and Y. Liu, "Big Data: A Survey", Mobile Networks and Applications, vol. 19, no. 2, 2014, pp. 171-209.
13. E. Bertino, "Big Data - Security and Privacy", 2015 IEEE International Congress on Big Data, 2015, pp. 757-761.
14. B. Matturdi, X. Zhou, S. Li, F. Lin, "Big Data security and privacy: A review", Big Data, Cloud & Mobile Computing, China Communications vol.11, issue: 14, pp. 135 – 145, 2014.
15. P. Adluru, S.S. Datla, Z. Xiaowen, "Hadoop eco system for big data security and privacy", Systems, Applications and Technology Conference (LISAT), Long Island, Farmingdale, NY, pp. 1 – 6, 2015.
16. S. Marchal, J. Xiuyan, R. State, T. Engel, "A Big Data Architecture for Large Scale Security Monitoring", Big Data (BigData Congress), pp. 56 – 63, Anchorage, AK, 2014.
17. T. Omer, P. Jules, "Big Data for All: Privacy and User Control in the Age of Analytics", Northwestern Journal of Technology and Intellectual Property, article 1, vol. 11, issue 5, 2013.
18. A. Kumar, L. HoonJae, R.P. Singh, "Efficient and secure Cloud storage for handling big data", Information Science and Service Science and Data Mining (ISSDM), pp. 162 – 166, Taipei, 2012.
19.] T. Vijey, A. Aiiad, "Big Data Security Issues Based on Quantum Cryptography and Privacy with Authentication for Mobile Data Center", Procedia Computer Science, vol. 50, pp. 149–156, 2015.
20. H. Chingfang, Z. Bing, Z. Maoyuan, "A novel group key transfer for big data security", Applied Mathematics and Computation, vol. 249, pp. 436–443, 2014.