Academic Guru Publication House

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

AG PH Books

A Review on Big Data: Security Challenges

A Sangeerani Devi[1*]

[1]*Assistant Professor, Department of Computer Science and Engineering , Sri Sairam Engineering College*

Abstract

Big data's widespread use has led to rapid growth of data resources, and new data analysis methods like conventional data mining and statistical analysis are helping to fuel this growth. In the world of big data, data from a variety of sources may be analysed, combined, and used in a variety of ways, resulting in new insights. Although each step of the life cycle offers data security and dependability concerns, the protection of personally identifiable information is a vital goal. Big data analytics, in particular, may be used to analyse user preferences, and this information can lead to the violation of personal privacy. The scope of big data is examined, as is the state of the art in big data security research. The concerns and causes influencing security are laid forth. The authors also touch on and expound on methods that protect individual privacy.

Keywords: large data; cycle of life; security of big data; privacy.

1. INTRODUCTION

There is a lot of interest in big data these days in business, science/technology/media and various government agencies. Healthcare, medicine, government agencies, distribution, marketing, and manufacturing are just a few of the industries used by many countries to turning big data in order to improve their services. It uses information-based technology to analyse vast amounts of data and predict future changes based on the information gained. Economic progress and technical advancement are both aided by this new source of energy. Big data is driven by a wide range of commercial and political objectives, including data integration, analysis, and mining. This is especially true when it comes to structured large data that comes from a wide range of sources, such as social media platforms, websites,

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

and global positioning systems. Data mining and statistical analysis approaches, such as standardised data mining, are propelling the growth of the market for big data because of their ability to mine large amounts of data quickly and efficiently. As a result of the data life cycle, which includes collection, analysis, fusion, and application, new information may be gleaned from big data.

2. ISSUES IN BIG DATA RELATED TO SECURITY AND PRIVACY

In light of Section II, discussion of the security and privacy implications of big data, this section takes a closer look at some of the most prevalent security and privacy concerns. There are a few conceptual and operational taxonomies of security and privacy provided by [1] and [2] to create vulnerabilities in big data systems. However, in order to establish a causal link between the properties of big data and vulnerabilities, we suggest and develop the following category based on the subjects and challenges in study disciplines, as shown in Fig.1:

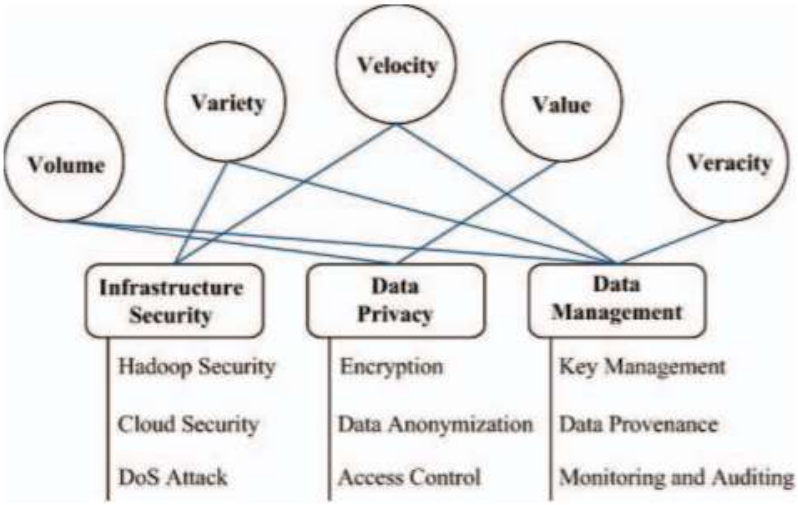

Figure 1. Category of Security Challenges in Big Data.

- Infrastructure Security
- Privacy of Data
- Data Management

There have been a lot of prior studies on the big data's security and privacy . The security and privacy issue may be better understood from several angles, as shown in Fig. 2.

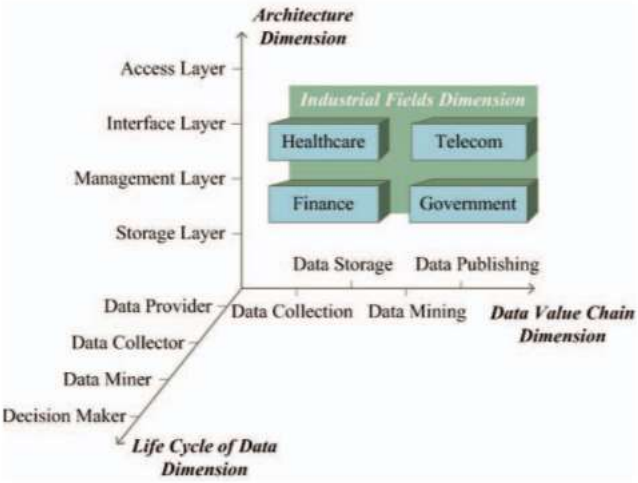

Figure. 2. Perspectives of Big Data Security and Privacy Analytics

3. ISSUES IN BIG DATA

Existing security for large data security and privacy are explored in this section. The researchers have devised a set of theoretical and operational privacy and security categories to better understand the threats presented by big data. At each of the four levels of a big data system, security and privacy are necessary, such as at the storage layer for the secure storage and control of monitoring devices. Hadoop Distributed File System (HDFS), data encryption, and so on [3] are included in the second management layer. There are three layers in the interface layer: an application programming interface (API), identity verification, and access control. Lastly, the access layer contains the user's cybersecurity. Big data presents a new set of challenges, as well as new opportunities, for organisations. This poll examines these challenges and opportunities. Security and privacy issues are examined and explored in detail. Furthermore, a study found that data sources, storage, and output all need some kind of security. The protection of the data types listed above will go a long way toward ensuring the safety of large amounts of data.

A variety of data encryption, access rights, transport layer security, and firewalls may be easily hacked. As a result of these factors, new technologies and strategies are being created to protect large data. The following portion of the study goes into depth into the Fig. 3 shown elements.

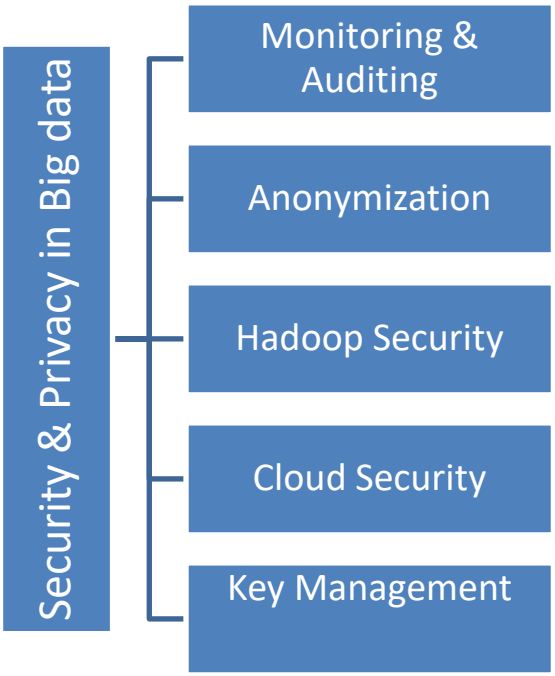

Fig. 3 Security and Privacy Categorization

A. Security in Hadoop

As one of the distributed process frameworks that isn't built for security, Hadoop is often used as an example. Big data analytics with Hadoop necessitates the use of a secure platform. In order to prevent hackers from stealing data from the cloud, two methods were recommended. HDFS has a trust mechanism that establishes a connection between the user and the name node. Authentication to the name node is required as part of the procedure. Users and name nodes create the hash function jointly. A comparison of the hash functions is made. If the two are identical, the user is granted access to the massive database. SHA-256, a hashing algorithm, is used for authentication in this method. The security of HDFS is of great importance. As a result, three strategies for enhancing security have been devised. The Kerberos system, which relies on a Service Ticket for security, is the first option. The algorithm known as the Bull Eye algorithm [4] is used to monitor sensor data and sensitive information in the second way. Replicated data and original data may be managed by this algorithm's primary benefit. HDFS security is further enhanced by using the Master Slave technique. There are two nodes in HDFS that act as a name node: the master and slave nodes. Using the Name Node Security Enhancement (NNSE) authorization, a slave node may respond and respond to an issue in the master node.

B. Monitoring as well as Auditing

For network security, detection of intrusions is an essential aim [5]. In order to identify intrusions on the whole network, DNS, HTTP, and other monitoring systems were created. Utilizing correlation

techniques, the scattered data is collected and analysed. To assess whether a node, flow, or packet has been malicious, the matrix has been set up. The detection system receives an alert message if any of these are detected or if the procedure is terminated by the prevention system. [5] Data availability, integrity, consistency, aggregation, and confidentiality are all factors that contribute to a security hole in huge data. There is a demand for security solutions to cover this gap in the market.

C. Security on Cloud Platform

Some of the reasons of cloud computing is extensively used because it provides on-demand services and resources that may be pooled. Despite the possibility of an attack, the cloud architecture's hosts are impervious to it. As a result, the cloud architecture service provider must take preventive measures. The cloud platform uses a variety of security measures, including authentication, compression, encryption, and decryption, to protect massive data. The cloud platform uses a security approach called Cryptographic Virtual Mapping to create a data route. Protection of vital, sensible and valuable pieces of information is the goal of this method. In order to protect sensitive, valuable, and vital data, the encryption is applied just to the storage channel that leads to it. Data parts and their accessing indexes must be available at all times in order to attain a factor of availability. This means that even if some data is lost, the overall availability of the system may be considered a success.

D. Anonymization

The volume characteristics of big data make it impossible for any of the traditional approaches to guarantee anonymity, despite the best efforts of researchers. In order to maximise scalability and privacy, a hybrid technique to anonymization has been developed that combines the traditional approaches of Bottom-Up and Top-Down. The t-ancestor clustering approach and a proximity-aware agglomerative algorithm are used to partition the dataset and record data, respectively, to overcome the issue of scalability [5]. Big data may be protected by using a differentiated privacy method. The model is constructed in such a manner that each piece of input data has an equal chance of being released. One of the two processes for guaranteeing differential privacy in large data is known as the Exponential mechanism and the Laplace mechanism. The Laplace mechanism is used to generate noise based on Laplace distribution for current and accurate results. The Exponential mechanism rewards outputs with higher scores with exponentially larger chances of success when the results are fictitious. This makes it more likely that you'll be referred to as [5]. Discriminating between private and public data is an important consideration in any research, even though it has been proved to be beneficial for trajectory data. As a result, it has been shown that all data sources, stored data, and output data need to be safeguarded [5]. The protection of the data types listed above will go a long way toward ensuring the safety of large amounts of data. In order to keep up with the ever-changing security landscape, methods like machine learning and statistical analysis, generally referred to as "data science," are often used. Analyses are performed depending on kind of the data present in the ecosystem, and a variety of machinery learning methods are used to spot any kind of alterations [6]. In order to safeguard big data environment, firms already utilise a variety of security techniques. Network device configuration may be managed using the IBM risk manager tool, which can be used to report and manage risks [7].

E. Key Management

A problem with large data security arises when keys are exchanged between the server and the user. A group key transfer mechanism is needed to distribute a key across numerous groups. As a result, a new protocol based on the Diffie-Hellman key agreement and a linear secret key method is implemented to guard against online attackers. As part of a collection of complicated networks, the Outsourcing Conditional Proxy Re-Encryption (CPRE) is used. Unstructured data, such as email, text, and XML, is also challenging to secure in large data systems [8]. Unstructured data may be protected by the use of data analytics techniques such as data filtering, grouping, and classification based on the degree of data sensitivity. In the second level, database's node of data is arranged, and the relevant and significant service (identity, integrity confidentiality, nonrepudiation, and authentication) is picked by a scheduling algorithm from the security suite to offer security.

4. CONCLUSION

Predicting future trends has become one of the most promising and dominant technologies because to the rise of big data. The privacy and the security should be taken into account when developing apps in these situations. When it comes to sensitive applications in big data, privacy and security must take precedence over anything else. This article examines the effects of big data features on infrastructure security and privacy, as well as cloud security and data management, on the privacy and the security of the big data systems. Data collection, storage, use, transport, and analysis all face unique challenges in the modern world. Security and privacy of Big data are our primary concerns. It's imperative that we create a framework for safe data sharing on a semi-trusted platform for big data that includes secure data transmission, storage, use, and destruction.

REFERENCES

1. Cloud Security Alliance Big Data Working Group, Expanded Top Ten Big Data Security and Privacy Chanllenges[R], Apr. 2013
2. NIST, NIST Big Data Interoperability Framework: Volume 4, Security and Privacy[R], National Institute for Standards and Technology, 2015, http://dx.doi.org/10.6028/NIST.SP.1500-4.
3. F. McSherry and K. Talwar, “Mechanism design via differential privacy,”in 48th Annual IEEE Symposium on Four.
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. 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.
6. 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.
7. 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.
8. 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.