RESEARCH ARTICLE OPEN ACCESS

Security and Privacy in Big data

SanthoshKumar K[1], Udhayakumar U[2]
1Department of Computer Science, Shanmuga Industries Arts and Science College, Tiruvannamalai
2 Department of Computer Science, Shanmuga Industries Arts and Science College, Tiruvannamalai

Abstract:

Big data is coming with new challenges in security; involve the three aspects of security: (confidentiality, availability, integrity) and privacy. These challenges are due to the characteristics 5V of data in Big data: velocity, variety, volume, value, and veracity. Moreover, it depends on several levels of security, network, data, application, and authentication. Furthermore, big data is also promising security. The massive amount of data provides more security information like data logs. Moreover, big data analysis can apply to security. Many theories for big data security are proposed in the literature, covering the different aspects of security and privacy. Recently, different schemes and frameworks are introduced to reach the high level of security in big data, based on different security theories. In this paper, we discuss different challenges in big data security and privacy, and we introduce recent security theories and works in this filed. A comparative study of latest advances in big data security and privacy presented.

Keywords — **Big data, Security, Privacy, Cryptography, access control.**

I. INTRODUCTION

Nowadays big data play an essential role in several areas of research and management such as internet transactions, social networks, retail trade, and healthcare. The vast amount of data generated by the enterprises and associations related to these domains cannot analyse with traditional approaches and applications [1]. For that, the attention of many types of research and managers has turned to big data analysis [3]. Big data analysis tools and applications aim to analyze data with the scalability and better performance. However, big data comes with new challenges concern security, which cannot cover with traditional security methods and tools. Challenges involve several levels: network, data, application, and authentication.

Level network: We outline first, different problems of network security in big data. The protocol SSL (Transport Layer Security) is one of the leading solutions which ensure network security that is used in general in the client-server model to establish a secure connection between client and server. SSL is a standard security technology for establishing an encrypted link between a server and a client [1, 2].

However, SSL consumes many processing resources on the server for encryption and decryption of secure data. To track this drawback of the client-server model, A load-balancing approach for scaling Secure Sockets Layer (SSL) proposed. One limitation of this approach that the load balancers cannot see any data, such as cookies or URL, inside the SSL traffic because the entire payload encrypted. This limits the load balancers in traffic-redirection and intelligent-switching capabilities for SSL traffic. For this purpose to attend high security with efficient processing in the network for big data, thinking for new architectures to avoid these problems is very interesting. Another point in network security is using Internet Protocol like IPv6 for network applications logging by using address IP, but this protocol is finitely scalable. Therefore, the need for a new architecture of network and computers that are infinitely scalable present a significant area of research for future big data security.

Level data: challenges at level data involve different components of information security. For integrity, detecting malicious data with the huge amount of data is very hard. Furthermore, the analysis of data which gathered from diverse sources in which one or more sources contain malicious data can affect results (unexpected results). Another point where integrity is not guaranteed when diverse sources of data over different nodes, failure of each node can affect the analysis of these data that can produce malicious results. As Big data contain a large number of user information in the data, this makes a big risk for big data privacy. For Big data confidentiality, volume and variety of data in which we have structured and unstructured data is a primary challenge for traditional cryptography approaches, for example, the encryption of spatial data that need great resources of computation [3].

Level application: main challenges at level application concerned integrity [4]. Detecting malicious codes in such applications running on Big data platform is very hard. A case of study is detecting instructed mappers in MapReduce [7,8]. Furthermore, a failure of the execution of such distributed network applications and services affect availability.

Level Authentication and access control: challenges belong to two categories: authentication and access control at the level application and authentication and access control at level data [6]. For the first, it is challenging to establish an access control scheme for applications running on Big data platform. A case of study: how to design efficient access control and authentication system for mappers and reducers in MapReduce. For the second, access to data by a high number of users make challenging to attribute and control privileges of this enormous number of users [9].

II. Recent Security Theories for Big Data

In literature, many security theories are introduced by researchers to cover the three aspects of big data security: (confidentiality, availability, integrity), and privacy, especially in the cloud environment. An intuitive way to achieve confidentiality protection for big data is to use encryption [10,11]. Classical encryption algorithms are used to ensure cryptography for general purpose processing (e.g., symmetric encryption algorithm AES and asymmetric encryption algorithm RSA). However, they are limited and cannot provide data processing. Therefore, the Homomorphic encryption technique is developed to solve this problem, which can achieve confidentiality protection and data processing at the same time. One drawback of these techniques of encryption is that they still need an expensive computation.

Other encryption algorithms like Classical Public Key Approach and Attribute-Based Encryption Approach proposed for data sharing. The main disadvantage of these algorithms is the high-cost computation on the owner side. To cover this limit authors in presented proxy re-encryption that is cheaper and flexible [12]. Proxy re-encryption assumes that the cloud is semi-trusted [5]. The main idea of this algorithm is as follows: first, data owner encrypts the data with a public key, after for each potential receiver, the data owner generates a re-encryption key. If such receiver is authorized to share data with the owner, then the cloud re-encrypt the cipher-text and sends it to the designed receiver for decryption.

Assuring integrity protection for big data is usually considered as the primary challenge. Classical digital signature in the environment of Big data is proposed, but it is not scalable, Untraceable, and based on strong assumption. Other solutions are introduced to tackle this issue such as HMAC /CMAC and homomorphic signature.

In HMAC/CMAC if one party in the system is involved, it will share secret keys with other parties, this leads to big risk if these parties reveal this secret key. Homomorphic signature depends on the security of nodes in the cloud system [13]. Privacy plays a more critical role in several fields like healthcare, and with the coming of big data, privacy protection becomes more critical. Differential privacy is a robust model that defined as data privacy protection. Moreover, has two properties that are used for big data, Sequential Composition, and Parallel Composition [14]. Recently, some works are done in the area of big data Security and Privacy to assure a high level of security for the two directions: Security and Privacy for big data and Big data for Security and Privacy, based on different security theories.

III. Big Data Security and Privacy

Big data comes with new challenges in security and privacy. However, big data also provides new trends for security and privacy; in this section, we present recent works in the two fields: Security and privacy for big data and big data for security and privacy.

A. Security and Privacy for big data

With the coming of big data, assuring security for the vast amount of data becomes a significant challenge in big data [15]. Traditional techniques and approaches are inadequate since they are designed to secure a small-scale data and don't respect the 5 V of Big data. Recently, some works and researchers focused on security and privacy for big data; they proposed different methods and solutions that meet different criteria of security: confidentiality, availability, and integrity, and privacy [3,15]. In authors proposed architecture of a hybrid cloud, composed of public cloud and private cloud. After data query from users, the private cloud store sensitive data after processing, and then send nonsensitive data to the public cloud [16]. This architecture aims to achieve image data privacy via the hybrid cloud, and reduce the time of computation by dividing the image into blocks and operate on these blocks, so that suitable for Big data.

An extension of big data technology framework is proposed by, that based on security architecture. This security architecture divided into the pre-filtering layer and the post-filtering layer. The pre-filtering layer is the first privacy layer of the proposed architecture [17]. It finds and deletes sensitive personal information from the collected data and stores them in the matching database system DB. The post-filtering layer filters and removes sensitive information synthesized after the significant data analysis and stores them in matching DB. Some other works focused on how to secure Sensitive Data Sharing on a Big Data Platform.

Authors in introduced a systematic framework for security sensitive data sharing on a big data platform. This framework is composed of three components: security submission, security storage, and security use. The primary flow of the framework is as follows: first personnel sensitive data are submitted to Big data platform using security plug-in [18]. After that, data stored in big data platform encrypted with Proxy-re-encryption, Then cloud platform service providers who want to share the sensitive information, download and decrypt the corresponding data from big data platform, in the private process space based on VMM using the secure plug-in. Finally, a secure mechanism is applied to destroy user data still stored temporarily in the cloud.

Another solution for securing Sensitive Data Sharing on a Big Data Platform is presented by authors. They proposed An Improved HABE Construction with Outsourced Decryption, which overcomes the limitations of the decryption cost at the user side that is still very high in traditional HABE. The main idea of the proposed HABE is decrypting the cipher text partially by the cloud after receiving the key outsourced from the user [19]. However, other researchers are interested in providing Access Control (AC) systems for big data. Authors in presented An Access Control Scheme for Big Data processing. The global structure of this scheme is as follows: MS (Master System) categorizes CS's (Cooperated System) by the security classes according to the SAs defined with BD source providers, And managed MSP that specifies a set of AC rules that are imposed by MS to enforce AC on CS's. CS managed CSP that allows the CS to control the access to the distributed BD data/process by considering the processing capabilities and security requirements of the CS. A FAD list defined as a federated dictionary of AC attributes that should be syntactically and semantically agreed by the MS and CSs [18, 19].

Recent researches developed different solutions for big data privacy based on Differential privacy which is a strong model for data privacy protection. A differential privacy protection scheme for big data in body sensor network is introduced by, based on the concept of dynamic noise thresholds.

The interference threshold is calculated for each data arrival to add noise to data. Furthermore, the same authors proposed another scheme based on differential privacy theory. Haar Wavelet transform method is used to convert histogram into a complete binary tree, for the purpose of reducing errors. Moreover, Big data models like MapReduce and its framework Hadoop are built without any secure assumption; new tools are developed to secure Hadoop such as Kerberos mechanism which is used to enhance the security in HDFS. Moreover, Apache Accumulo that allows multi-level access control at the cell level in a key-value store.

B. Big data for Security and Privacy

Big data comes with new trends for security and privacy. Big data analytics offer a large scale analysis and processing a huge amount of structured an uninstructed data in big data, for that becomes an important research area in security [20]. Moreover, big data provide a great amount of information like data logs. These logs are gathered from traffic network, applications, data, and users. So that searching for a correlation between these logs can be a key for detecting malicious codes and activities in security.

Many research works focused on analysing security log system using big data. The proposed methods to filter and analyse logs system. In authors proposed intelligent information analysis platform for system construction of security log analysis using Big data which is composed by collecting, saving, processing, and analysing techniques. This architecture aims to analyse the relationship between security and data events created from network, system, application service of central IT infrastructure. In Other work, authors outline that the user behaviours are dynamic which is challenging to capture the users' comprehensive behaviours in a single device by capturing or collecting the static dataset [17, 20]. Therefore, they proposed a log analysis system which is based on the Hadoop distribution platform to capture the traffic and analyse the user & machine behaviours, in terms of the search keywords, user shopping trends, website posts and replies, and web visited history to acquire the uses' dynamic behaviours.

Big data turns the way on the architecture and design of networks. Traditional networks are inadequate for managing and processing big data and don't respect the 5V of big data. They are composed of many layers and need an expensive computation over these layers. Therefore, new technologies have emerged. Recently, software-defined networking (SDN) has attracted considerable interest as a new paradigm in networking. It is composed of fewer layers which make it more flexible and efficient for big data.

The main idea of SDN is to detach the control plane from the forwarding plane, to break vertical integration, and to introduce the ability to program the network. So that from a security point of view the 5 V's of Big Data demands ultra-fast response times from security and privacy solutions/products.

Furthermore, providing integrity for security implies not only detecting malicious code and data, but also intervene at the right moment, due to the flexibility of this new architecture of the network. For this purpose, SDN is a suitable paradigm in networking that benefits security and privacy in big data.

IV. RECENT TRENDS IN BIG DATA SECURITY AND PRIVACY

We highlight a study of different researchers in the recent trends in the area of big data security and privacy.

A. *PROTECTING TRANSACTION LOGS AND DATA*

Data stored in storage medium, such as transaction logs and other sensitive information, may have varying levels, but that's not enough. For instance, the transfer of data between these levels gives the IT manager insight over the data which is being moved. Data size being continuously increased the scalability and availability makes auto-tiering necessary for big data storage management. Yet, new challenges are being posed to big data storage as the auto-tiering method doesn't keep track of data storage location.

B. *VALIDATION AND FILTRATION OF END-POINT INPUTS*

End-point devices are the main factors for maintaining big data. Storage, processing and other necessary tasks are performed with the help of input data, which is provided by end-points. Therefore, an organization should make sure to use authentic and legitimate end-point devices.

C. *SECURING DISTRIBUTED FRAMEWORK CALCULATIONS AND OTHER PROCESSES*

Computational security and other digital assets in a distributed framework like MapReduce function of Hadoop, mostly lack security protections. The two main preventions for it are securing the mappers and protecting the data in the presence of an unauthorized mapper.

D. *SECURING AND PROTECTING DATA IN REAL TIME*

Due to large amounts of data generation, most organizations are unable to maintain regular checks. However, it is most beneficial to perform security checks and observation in real time or almost in real time.

E. *PROTECTING ACCESS CONTROL METHOD COMMUNICATION AND ENCRYPTION*

A secured data storage device is an intelligent step in order to protect the data. Yet, because most often data storage devices are vulnerable, it is necessary to encrypt the access control methods as well.

V. CONCLUSIONS

In this paper, the main challenges in the area of big data security and privacy introduced. We also discuss different security theories for big data. Moreover, a comparative study of recent researches on big data security and privacy presented in which we outline for each work the information security criteria achieved and different advantages and limits of the solution proposed by this work. We conclude that recent works on big data security are improved but still suffer from such limits that decrease their performance and designed for a particular purpose. Therefore, development of abstract and unified models and frameworks for big data security that ensure all criteria of security is very interest. In future, we aim to propose and design an optimization framework for big data security analysis.

ACKNOWLEDGMENT

I would like to convey my heartfelt thanks to Mr. Udhayakumar, my research guidance for completion of my project. He helped me to understand and remember important details of the project. He helped me and gave his guidance in completing of my project successfully.

REFERENCES

1. *Akoka, J., Comyn-Wattiau, I., & Laoufi, N. (2017). Research on Big Data–A systematic mapping study. Computer Standards & Interfaces, 54, 105-115.*

2. *Bechini, A., Marcelloni, F., & Segatori, A. (2016). A MapReduce solution for associative classification of big data. Information Sciences, 332, 33-55.*

3. *Berger, M.L., & Doban, V. (2014). Big data, advanced analytics and the future of comparative effectiveness research. Journal of comparative effectiveness research, 3 2, 167-76.*

4. *Big Data for dummies – Judith Hurwitz et al., Wiley, 2013.*

5. *Big Data, Big Analytics – Michael Minelli et al., Wiley, 2013.*

6. *Bradlow, E. T., Gangwar, M., Kopalle, P., & Voleti, S. (2017). The role of big data and predictive analytics in retailing. Journal of Retailing, 93(1), 79-95.*

7. *Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. Mobile Networks and Applications, 19(2), 171-209.*

8. *Choi, T. M., Chan, H. K., & Yue, X. (2017). Recent development in big data analytics for business operations and risk management. IEEE transactions on cybernetics, 47(1), 81-92.*

9. *Ducange, P., Pecori, R., & Mezzina, P. (2017). A glimpse on big data analytics in the framework of marketing strategies. Soft Computing, 1-18.*

10. *Elgendy, N. and Elragal, A., 2014, July. Big data analytics: a literature review paper. In Industrial Conference on Data Mining (pp. 214-227). Springer, Cham.*

11. *D. Ranjith, J. M. Balajee and C. Kumar," Trust computation methods in mobile ADHOC network using glomosim: A Review" International Journal of Scientific Research and Modern Education, Vol. I, Issue I, pp. 777-780, Nov.2016.*

12. *Janarthanan Y, Balajee J.M, and Srinivasa Raghava S. "Content based video retrieval and analysis using image processing: A review."International Journal of Pharmacy and Technology 8, no.4 (2016): 5042-5048.*

13. *Jeyakumar, Balajee, MA Saleem Durai, and Daphne Lopez. "Case Studies in Amalgamation of Deep Learning and Big Data." In HCI Challenges and Privacy Preservation in Big Data Security, pp. 159-174. IGI Global, 2018.*

14. *Kamalakannan, S. "G., Balajee, J., Srinivasa Raghavan.,"Superior content-based video retrieval system according to query image"." International Journal of Applied Engineering Research 10, no. 3 (2015): 7951-7957.*

15. *Priya, V., Subha, S., & Balamurugan, B. (2017). Analysis of performance measures to handle medical E-commerce shopping cart abandonment in cloud. Informatics in Medicine Unlocked.*

16. *Rangith. D Lakshmi narayanan. J and Balajee. J," A study of behavior on information system in a university campus by analysis of people mobility" International journal of research in computer application & management, Vol. 6, Issue 7, pp. 29-31, Jul.2016.*

17. *Ranjith, D., J. Balajee, and C. Kumar. "In premises of cloud computing and models." International Journal of Pharmacy and Technology 8, no. 3 (2016): 4685-4695.*

18. *Sethumadahavi R Balajee J "Big Data Deep Learning in Healthcare for Electronic Health Records," International Scientific Research Organization Journal, vol. 2, Issue 2, pp. 31–35, Jul. 2017.*

19. *Ushapreethi P, Balajee Jeyakumar and BalaKrishnan P, Action Recongnition in Video Survillance Using Hipi and Map Reducing Model, International Journal of Mechanical Engineering and Technology 8(11), 2017,pp. 368–375.*

20. *Zhang, W., Zhao, Q., Deng, J., Hu, Y., Wang, Y., & Ouyang, D. (2017). Big data analysis of global advances in pharmaceutics and drug delivery 1980–2014. Drug Discovery Today.*