International Journal of Engineering & Technology, 7 (4.39) (2018) 734-736

SPC

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Survey on Data Security and Privacy Preserving in Big Data

A.Vineela *, N. Kasiviswanath [2], L. Sudha Rani

[1] *Research Scholar, Dept. of CSE, JNTU Anantapuram, Andhra Pradesh, India*
[2] *Professor & HoD, Dept. of CSE, GPREC, Kurnool, Andhra Pradesh, India*
[3] *Assistant Professor, Dept. of CSE, GPREC, Kurnool, Andhra Pradesh, India*
**Corresponding author E-mail: vineela177a@gmail.com*

Abstract

With the rapid growth of IT industry, big data needs the improvement in storage, computation and network field. This enhancement also brings the new security and privacy issues to the big data. The researchers are attracted towards to solve the security and privacy issues. This paper made a survey on characteristics of big data along with security issues. The traditional security methods of cloud computing are not appropriate to the big data. Privacy preserving is also one major issue in big data. This survey also provides complete study on research issues and challenges of privacy preserving and the comparison is made to the privacy preserving techniques. Finally, this paper provides comprehensive overview of the methods to solve the big data security problem.

Keywords: *Big Data; Security; Privacy; Data encryption; Data mining.*

1. Introduction

In the recent years, big data is the buzzword all over the internet. The word big data denotes to the large volume of data generated by the internet users and the IT industries. Due to the inventions of cloud computing and big data leads to rapid growth in the utilization of social networks and information technology [2-3]. Based on the recent survey, there were 200 thousand users searching the content in the Google per second and the 40 billion Facebook users are sharing the digital content daily. The social networks have continuous producing the huge volumes of data and also the fields like medicine, finance, scientific computing and retails are generating the variety of data by using the sensor and monitoring devices. The increased volumes of data need efficient computing techniques to process the data. Therefore, big data received great attention by the researchers and industries.

Big data is treated as the new interesting environment that motivates the technology and business inventions, as well as financial growth. The data integration, data analysis and data mining are the techniques that are involved in the data processing of big data. The major issue to deal with the big data is data security and privacy. Every activity performed by the users in the Internet is known to the internet providers. For instance, the Flipkart knows our shopping habits, Google is aware of our searching habits, Facebook knows our friends list. Due to this, the cloud and big data fields are facing many challenges in providing privacy and security. It has the connection with technology, ethics, morality, management and commercial interests. Therefore, providing big data security and preserving the privacy is more difficult than the conventional security issues [1]. The implementation of data security and privacy involves two considerations. First one is to ensure the data privacy at the application level and the second one is to provide the security at the time of application usage.

Based on the applications of big data, this research work concentrates on latest improvements in data security and privacy preserving techniques. This study concentrates on special aspects such as encryption, access control, data auditing, anonymous protection and different security analysis. This research work list some issues and challenges to guide the academicians and researchers to continue their further research. The rest of the paper is organized as follows. Section 2 deals with the classifications of Big Data. Section 3 explains about security issues in Big Data. Section 4 deals with privacy preserving techniques in big data and finally the conclusion is drawn in Section 5.

2. Classification of Big Data

In general, big data is often represented with 3V's such as Volume, Velocity and Variety, as shown in Fig. 1. Volume represents the massive amount of data, i.e, it contains huge volumes and data sets and it requires huge computation for analysis. The volume of the data is ranges from TB to PB. Velocity represents the speed of the computation. It needs to perform the data computation in lightning speed. For instance, the video monitoring system continuously monitors the data and identifies the useful data in matter of seconds. This mechanism is different from traditional data mining approaches. The variety represents the categories of data. The data sets in the big data are collected from different sources are of different formats such as unstructured, structured and semi-structured. The definition of big data is given as follows:

"Big data is defined as the data set whose computation time is more than the tolerable time in using the traditional software tools to store, manage and manipulate the data".

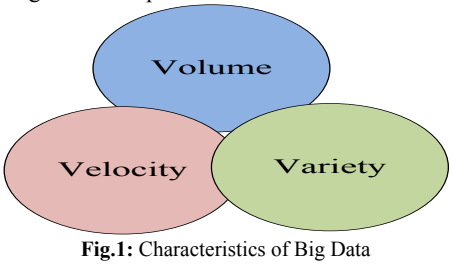


Fig.1: Characteristics of Big Data

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

2.1. Framework of Big Data

Based on the complete analysis of data processing, big data not only involves the process of data storage, data pre-processing, data acquisition and management but also it contains data analysis, data mining, virtualization, data security and data privacy [4-6]. Figure 2 shows the schematic representation of big data framework [7]. This survey only focuses on the security and privacy issues in all the modules of big data analytics.

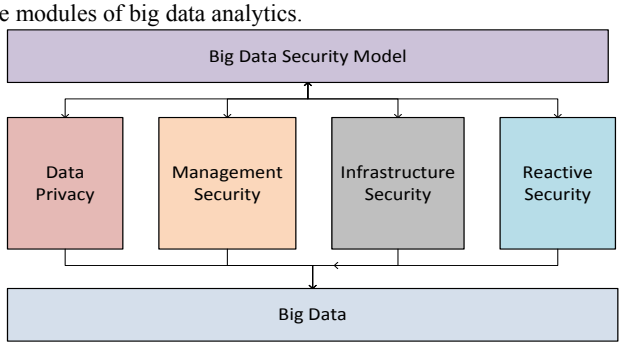


Fig.2: Semantic Representation of Big data Framework

3. Security Issues in Big Data

According to the Gartner [8], security in big data is a major issue to deal with. Today, many industries are penetrating towards the big data. Due to having the high processing capacity and analytical technology, big data will capture the important data in order to make the decision based on the client's request. Apart from the big data benefits, some of the research challenges are given in Fig. 3.

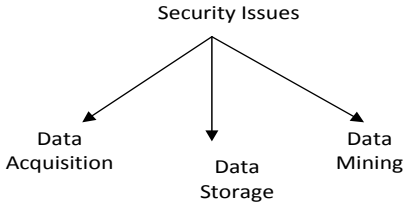


Fig. 3: Security Issues in Big Data

3.1. Data Acquisition

Diversity is the major source for big data. As an initial step in big data processing, the data has to be collected and pre-processed to retrieve the useful information. Big data not only contains the large volumes of data but also contains the sensitive and complex data [9]. Therefore, the data is more attracted by the attackers and also the data integration provides chances to the attackers to employ successful attacks. The confidentiality in big data refers to the providing restriction to the unauthorized users, processes or entities. A huge volume of data of data contains more number of organizations, personal and all kinds of records. The storage of these records in to central data base leads to the data leakage.
Data integrity is one of the factors which influence the big data. In general, data integrity refers to the process of providing access to the data by the authorized people. The major motive of data integration is to restrict the unauthorized access. Due to the sloppiness in the big data network structure, the data would be damaged by tampering, forgery, interruption and interception. The encryption mechanism is the technique that can solve the above discussed issues, but it not completely fit for all the security problems.

3.2. Data Storage

The forum of network created the opportunity to share the resources and to exchange the data in big data processing. In the recent years, many user accounts have been stolen in the internet. This is happen due to the compromise of big data by the attackers. Before the big data has been evolved, the data is stored in the form of file servers and relational data base formats. For unstructured data, the data is stored in the NoSQL format. This format is advantageous in availability and scalability and act as substitute to the big data. Apart from the advantageous, the NoSQL is lack in privacy management and access control.

3.3. Data Mining

With the advancement of computer networks and artificial intelligence, the data mining approaches are widely used for data collecting and analysis. Apart from this, big data is major carrier for different attacks. The data sets may contain malicious software that is hard to find by the traditional intrusion detection systems. The attackers compromise the big data in the following aspects. The investigations in [10-12] proved that the failure in handling the big data causes major threats to user privacy. Based on the diversity in the data, the privacy is divided in to location aware privacy and anonymous identification. The attacks faced by the users are not only on the personal data but also on the behaviour of the users on the big data.

4. Onion Model for Big Data Security

Big Data is one of the fields that have more benefits to the enterprises, by uncovering the customer buying habits, monitoring the real time events and identifying and preventing the frauds. However, the applications in big data with poor security lead to data breaches. Therefore, Big Data needs to be protected by ensuring the authorized people access permits. The security in Big Data needs to address several mechanisms for large computing resources and large forms of data. As the computing infrastructure and data size is more, the traditional computing mechanisms are failed to scale and secure the data and also the hybrid cloud infrastructures provides the advantages to the attackers to gain the access over the network. The onion model of the Big Data security is given in Fig. 4.

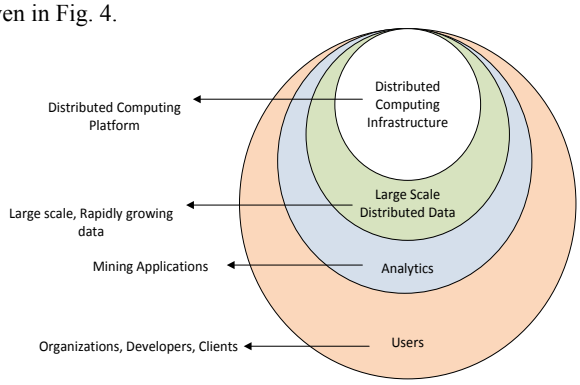


Fig. 4: Onion Model for Big Data Security

Fig.5 describes about several security concepts in Big Data. The detailed explanation about each security concept is given below:

User security and Privacy: Authentication, Integrity and Confidentiality methods to validate the users.
Analytics Security: Providing security to the big data applications and managing the proper analytical tools to the clients and end users.
Large Scale Distributed Data: Applying the encryption mechanism and privacy preserving mechanism to secure the stored data on Big Data environment.

Distributed Computing Infrastructure: Managing security over multiple distributed environments for data analysis process.

Security in Big Data refers to both data security and computation security, preserving the data through the life span, i.e. from crea-

tion of the data to the disposal of the data. The data must be secure while in transit, processing and idle. During the time of data processing, the data must be passed with different stages. Therefore, managing the security at different levels is very important. Big data Security is majorly classified into three categories. Storage and transmission data security, privacy preserving and user level security. Figure 4 shows the security components in Big Data.

4.1. User Level Security

Authentication, integrity, authorization, confidentiality mechanism for data access and computing to be managed based on the privileges granted to the users, such as user authentication is performed in automated fashion. The improper management of user authentication leads to security breaches and it proves the non-compliance of the organizations in audit. Therefore, a proper user level security is needed to preserve the data from the intruders.

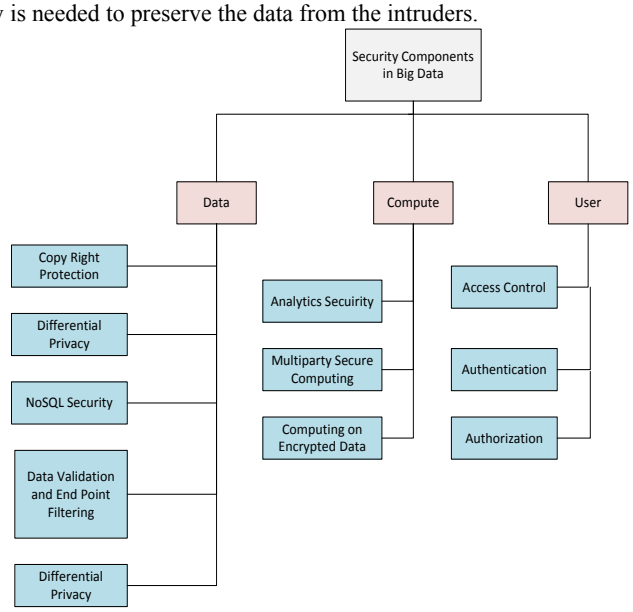


Fig. 5: Security Components in Big Data

4.2. Storage and Transmission Data Security

Secure communication plays a crucial role in user authentication and data encryption over the network. The data encryption mechanisms are broadly addressed by the cryptographic mechanisms. In general, the sensitive data is stored in the cloud infrastructures without encryption mechanisms. The major issue to perform encryption is due to large volumes of data which will not allow the users easily to search and share the records. Security mechanism such as attribute based encryption lessens the problem by introducing the public key cryptography for accessing the stored data. The transmission data is secured by applying the cryptographic enforced secure communication. Table 1shows the security elements of Big Data.

Table 1: Security elements in Big Data

Security elements	**Characteristics**
Authentication and Authorization	Validating the user identity and granting the permissions.
Availability	The data should be avail when it is needed by the user.
Data Base and File Security	Access permissions based on the user roles in the Big Data.
Secure Computing	Preventing exposing of computation to the other parties.
Analytics Security	Preventing unauthorized access to applications thus retaining integrity.
NoSQL Security	NoSQL databases come with little built-in security. They have what's called BASE (Basically Available, Soft state, Eventually consistent) properties; rather than requiring consistency after every transaction, the data base just needs to eventually reach a consistent state.
Copy Right Protection	Protecting the data from tampering and copying by the intruders.
Data validation and filtering	Identifying the data which is required and reduce the expenditure and computation cost

5. Conclusion

Big data is treated as the new interesting environment that motivates the technology and business inventions, as well as financial growth. The major issue to deal with the big data is data security and privacy. This paper defined the big data concepts along with security and privacy challenges. This survey shows that the existing security and privacy methods are not enough to support the big data. The conventional privacy preserving models includes storage encryption, data perturbation, identity authentication, secure computing and access control etc. But, single privacy preserving model is not sufficient to all the applications in big data. The best solution is to integrate the different methods and related policies based on the user requirements. Then, the security and privacy will be achieved better.

References

[1] Kim, H. (2013). Privacy preserving security framework for cognitive radio networks. *IETE Technical Review*, *30*(2), 142-148.
[2] Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity.
[3] J. Li, and X. Q. Cheng (2012), "Big data research: Major strategic field for future science and technology and the development of the economic society," Bull. Chin. Acad. Sci.,27, (6), 647-57.
[4] Feng, D., Zhang, M., & Li, H. (2014). Big data security and privacy protection. *Chinese Journal of Computers*, *37*(1), 246-258.
[5] Xiaofeng, M., & Xiang, C. (2013). Big data management: concepts, techniques and challenges [J]. *Journal of computer research and development*, *1*(98), 146-169.
[6] Kshetri, N. (2014). The emerging role of Big Data in key development issues: Opportunities, challenges, and concerns. *Big Data & Society*, *1*(2), 2053951714564227.
[7] Fang, W., Zheng, Y., & Xiu, J. (2014). Big data: Conceptions, key technologies and application. *Nanjing Xinxi Gongcheng Daxue Xuebao*, *6*(5), 405.
[8] Kim, G. H., Trimi, S., & Chung, J. H. (2014). Big-data applications in the government sector. *Communications of the ACM*, *57*(3), 78-85.
[9] Jagadish, H. V., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J. M., Ramakrishnan, R., & Shahabi, C. (2014). Big data and its technical challenges. *Communications of the ACM*, *57*(7), 86-94.
[10] Sweeney, L. (2002). k-anonymity: a model for protecting privacy'International Journal on Uncertainty, Fuzziness and Knowledge-based Systems 10, 5 (2002) 557-570.
[11] Li, N., Li, T., & Venkatasubramanian, S. (2007, April). t-closeness: Privacy beyond k-anonymity and l-diversity. In *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on* (pp. 106-115). IEEE.
[12] Ying, X., & Wu, X. (2008, April). Randomizing social networks: a spectrum preserving approach. In *proceedings of the 2008 SIAM International Conference on Data Mining* (pp. 739-750). Society for Industrial and Applied Mathematics.