PROTECTING UNAUTHORIZED BIG DATA ANALYSIS USING ATTRIBUTE (DATA) RELATIONSHIP

Shaukat Ali[1*], Azhar Rauf[2], Jamil Ahmad[1]
[1]Department of Computer Science, Islamia College Peshawar (Chartered University), Pakistan
[2]Department of Computer Science, University of Peshawar, Pakistan
[*]Corresponding Author: shaukat@icp.edu.pk

ABSTRACT: *The term big data is a buzz word and hot area of research in the research community and industry. The term big data is not a new term, it the extension or new version of old terms, for example, the "Information Explosion". Security of big data is a challenge for researchers. It is difficult to protect all data in the big data environment, but it will be better to protect the value which can be extracted from big data. Value is the information extracted from data, and it is more important than the data itself. In this paper the value of data is protected instead of data. If we store the data in such a way that only authorized users are allowed to analyze it and* extract information from it. The unauthorized users will be unable to extract information or value from it.

Keywords: Big Data, Big Data Security, Attribute Relationship

I. INTRODUCTION

Several new technologies have been developed including development of network technology and distributed computing environment. With the introduction of new IT technologies and the Internet, a large amount of digital data is produced and distributed on daily bases. As a result, data analytic process has been accelerated to deal with large volume of information [1]. The analysis of such big data may lead to security leakage.

Big data security is a hot topic for research in today's research community and IT industry. Big data itself is a buzz word. Big data analytics is important for all those organizations who can extract value from it. The utilization of big data can able the organizations to lead the competitors. The value is the summarized information extracted from large amount of data. This value is important with security point of view. In the study of big data, the data is compared to minerals. It means that value extracted through the analysis is more important than the data itself [2]. The security of big data is considered to be important because it has important information hidden in it. In recent years, an active research is on going in the big data security through authentication or access control mechanism etc. [3]

This paper focuses on the security of the value extracted from the big data. Since the information (value) is more important than data, therefore, the security of value is more important than the security of data.

II. CONCEPT OF BIG DATA

Big data is buzz word in today's research community. The term "Big Data" was introduced by Roger Magoulas O'Reilly media for the first time in 2005 [4]. The definition of big data itself is a debatable topic. Different researchers define the big data in different ways. Yan et al. [5] defines big data as data that is huge, and has heterogeneous data and involves complex data manipulation. IDC defines big data as ''Big data technologies describe a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling high-velocity capture, discovery, and/or analysis''[6] The Gartner defines Big Data as follow with three parts [7, 8]: "Big data is high-volume, high velocity and high-variety information assets that demand cost effective, innovative forms of information processing for enhanced insight and decision making." Big data refers to datasets or combinations of datasets whose size (volume), complexity (variability), and growth rate (velocity) make it difficult to be gathered, managed, or analyzed by traditional technologies and tools, such as relational databases and statistics or visualization tools, within the time necessary to make them useful [9]. Definition by Jason Bloomberg is more simple [10]: "Big Data: a massive volume of both structured and unstructured data that is so large that it's difficult to process using traditional database and software techniques."Some of the researchers consider the size of big data as a part of problem and such data is big data [11]. The big data can be differ from regular data using the following four V's [12].

Volume: Petabytes of data have been generated on daily basis from different resources. Social media generates a large amount of data which is user generated contents. Machine generated data is also has a huge volume. Volume of data is a key component of big data.

Velocity: According to IDC the entire volume of data in 2006 was 0.8 Zetabytes, but in 2020 the digital data will be 40 Zetabytes [13]. The speed of digital data generation is quite high. Only Facebook generate 25 terabytes of data per day. The volume of big data increases with a high speed.

Varity: Data within big data has different formats. It has structured, unstructured, and semi-structured data. Big data includes sensor data, market campaign data, relational and non-relational data etc. The Varity component of big data makes it difficult to manage such data.

Value: Value is an important component of big data. Value is the information hidden inside the big data. Value is more important as compared to data. In this paper the value component is considered for security point of view.

A. The importance of Big Data

One the high importance of big data is the value inside the data. Value is the information hidden within the big data. This research paper focuses on the security of value instead of data. The organizations can utilize big data to lead their competitors in different areas. The big data can efficiently be utilized the following areas [4].

- In the fraud detection in the online transactions for any organization
- In risk assessment by analyzing data from the transactions in financial organizations.
- In the IT environment to easily troubleshoot the problem by analyzing log information.
- While improving the customer services by analyzing the blogs and social media.

III. RELATED WORK

Very little work has been done in the area of big data security although an active research is going on in this area. In the big data environment security is considered as secondary issue till yet. Many of the big data and NoSQL tools are developed with keeping the problem of data storage and security was not the primary concern in NoSQL databases [14]. Some of the researchers consider big data as an extension of early data storage methods like data warehouse [11,15] but the traditional security mechanism cannot provide better security measures for big data.

IV. CONTRIBUTION

In today's digital world, information is money, which comes from data. It can generate value if the right information is extracted to the right person at right time from the digital universe [6]. The value generated from data is more important than the data itself, therefore, it is focused in this paper to protect the value. The value or information can be protected by protecting the relationship between the data. If the relationship between the data (attributes) is hidden, no information can be generated from that data. Technically it is difficult to protect all the relationships between the entire data. It is, therefore, proposed in this research work to protect the selected relationships. Only the important relationship will be selected for protection. The importance of the relationship can be calculated from the importance of attribute. The attribute which has relationship with more attributes can be considered as an important attribute and the relationship of same attribute will considered be important relationship. For example, consider the table (Table I), which is the original table containing the entire data having sensitive and non-sensitive data. In order to secure the sensitive data, the table is split into two tables. One table (Table II) contains non-sensitive data and not needed to secure it. The other table (Table III) containing the sensitive data. If we want to secure it, its relationship with other data will be hidden in order to prevent unauthorized users to extract values from it. If we secure the value of data it means the data is secured. To hide the relationship of data from other data, the sensitive data will be kept separate from other data along with the key (Primary key) attribute but the key attribute will be secured using any technique. Encryption is one of the techniques to secure the key attribute. The same protection can be gotten by using the protection of attribute itself while using encryption, but the encryption of actual attribute will reduce the performance of analysis. In the analysis process the time is important because the information is needed at right time to right person. If the information is not available at right time it is useless.

Table I: Original table containing sensitive and non-sensitive data.

Key	Attribute1	Attribute2	Attribute3
1	Text data	Sensitive data	Numeric data
2	Text data	Sensitive data	Numeric data
3	Text data	Sensitive data	Numeric data
4	Text data	Sensitive data	Numeric data
5	Text data	Sensitive data	Numeric data

Table II: Table containing non-sensitive data only

key	Attribute1	Attribute3
1	Text data	Text data
2	Text data	Text data
3	Text data	Text data
4	Text data	Text data
5	Text data	Text data

Table III: Table containing sensitive data only

Key	Attribute2
Encrypted	Numeric data
Encrypted	Numeric data
Encrypted	Numeric data
Encrypted	Numeric data
Encrypted	Numeric data

V. CONCLUSION AND FUTURE WORK

In this paper the concept of big data security is discussed. A novel concept of big data security is presented to protect the value of data. Value of data is the useful information extracted from huge amount of data using analysis techniques. The value can be protected by hiding the relationship between the data.

REFERENCE

[1] J. Bughin, M. Chui, and J. Manyika, "Clouds, big data, and smart assets: Ten tech-enabled business trends to watch," *McKinsey Quarterly*, vol. 56, pp. 75-86, 2010.

[2] S.-H. Kim, J.-H. Eom, and T.-M. Chung, "Big Data Security Hardening Methodology Using Attributes Relationship," in *International Conference on Information Science and Applications (ICISA)*, Suwon, Korea, 2013, pp. 1-2.

[3] C. Tankard, "Big data security," *Network security*, vol. 2012, pp. 5-8, 2012.

[4] E. G. Ularu, F. C. Puican, A. Apostu, and M. Velicanu, "Perspectives on Big Data and Big Data Analytics," *Database Systems Journal*, vol. 3, pp. 3-14, 2012.

[5] K. W. Yan, N. M. Perumal, and T. Dillon, "Data migration ecosystem for big data invited paper," in *7th IEEE International Conference on Digital Ecosystems and Technologies (DEST)*, 2013, pp. 189-194.

[6] J. Gantz and D. Reinsel, "Extracting value from chaos," 2011.

[7] Gartner, "Big Data defintion, Gartner, Inc. [Online]. Available: http://www.gartner.com/it-glossary/big-data."

[8] S.Sicular, "Gartner's Big Data Definition Consists of Three Parts, Not to Be Confused with Three "V"s, Gartner, Inc. 27 March 2013. [Online]. Available: http://www.forbes.com/sites/gartnergroup/2013/03/27/ gartners-big-data-definition-consists-of-three-parts-not-to-be-confusedwith-three-vs," 2013.

[9] "Why is BIG Data Important? A Navint Partners White Paper " 2012.

[10] J. Bloomberg, "The Big Data Long Tail. Blog post by Jason Bloomberg, January 17, 2013. [Online]. Available: http://www.devx.com/blog/the-big-data-longtail.html," 2013.

[11] A. Lane, "Securing big data-Security recommendations for hadoop and NoSql environment," Securosis, 2012.

[12] J. P. Dijcks, "Oracle: Big Data for the Enterprise. Oracle Corporation," 2012.

[13] J. Gantz and D. Reinsel, "The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east," 2012.

[14] L. Okman, N. Gal-Oz, Y. Gonen, E. Gudes, and J. Abramov, "Security issues in nosql databases," in *IEEE 10th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, 2011, pp. 541-547.

[15] S.-H. Kim, N.-U. Kim, and T.-M. Chung, "Attribute Relationship Evaluation Methodology for Big Data Security," in *IEEE International Conference on IT Convergence and Security (ICITCS)*, , 2013, pp. 1-4.