IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol.-6, Issue-5, May 2018 E-ISSN: 2347-2693

Big Data and its Security Issues

B. Duhan[1*], D. Singh[2]

[1] Dept. of Computer Science and Engineering, DCRUST, Murthal, Sonepat, India
[2]Dept. of Computer Science and Engineering, DCRUST, Murthal, Sonepat, India

[*]*Corresponding Author: bhartiduhan12@gmail.com, Tel.: 7876872111*

Available online at: www.ijcseonline.org

Accepted: 16/May/2018, Published: 31/May/2018

***Abstract*—** Big data, as the term implies is a collection of huge amount of data. This can be valuable information to any organization. This huge amount of data requires emerging new technologies and design that makes it easier to take out information. As today world is connected virtually through internet so daily data generated is very high like in the form of zettabytes. As complexity of data has increased, there is a difficulty in managing such data by using traditional computing resources. It's necessary to secure big data environments from the infringement of confidential data. This paper presents introduction of big data and various security issues in this area.

***Keywords*—**Big data, Security, Privacy

I. INTRODUCTION

Big Data, as the term implies is used for describing high volume of data. Nowadays data grows at a very high velocity so it is tough to manage this huge amount of data. This data can be in exabytes, zettabytes or even more than that. There is a difficulty in managing this much data because volume increases at a fast rate as compared to traditional computing resources. Big Data mainly concentrates on quantity of data rather than quality of data. Data can be both structured as well as unstructured. Amount of data is not important but how organization uses that data is important [4]. As its name points that it is related to size of the data only but ignores other existing properties. It is necessary to secure big data environments from the attacks and protects leakage of confidential data. These attacks includes spamming attacks, Search Poisoning, Botnets, Denial of service attack (DOS), Phishing, Malware, website threats. To provide security from attacks all other properties of Big Data must be known.

Mainly Big Data can be described by 3 Characteristics [5].These are known as 3 V's:-

1.1 *Volume* (Data in Rest):-The quantity of data which is generated is important in this area. Size determines the potential and value of data which has to be considered. The name 'Big Data' itself shows the importance of this characteristic. In short, volume means data in rest. Data can be collected from different sources such as social media, commercial transactions, and information from sensors etc.

1.2 *Velocity* (Data in Motion):-In this context the velocity refers to the data generation's speed which means at what rate the data is generated. This data is being processed to meet the demands which are required for growth and development. Data is generated at a different unmatched speed from same or different sources of data. Different type of IOT sensors, RFID tags are used to tackle flow of data in real time scenarios.

1.3 *Variety* (Data in different Forms):-Data can be categorized in many categories .It can be numeric data, unstructured text document, structured document, audio, video, emails, stock transactions and financial transactions. The category of data is a important fact which has to be uncovered by data Analysts. Data Analysts can analyse data better by knowing its category. These 3 V's are explored in figure shown below.

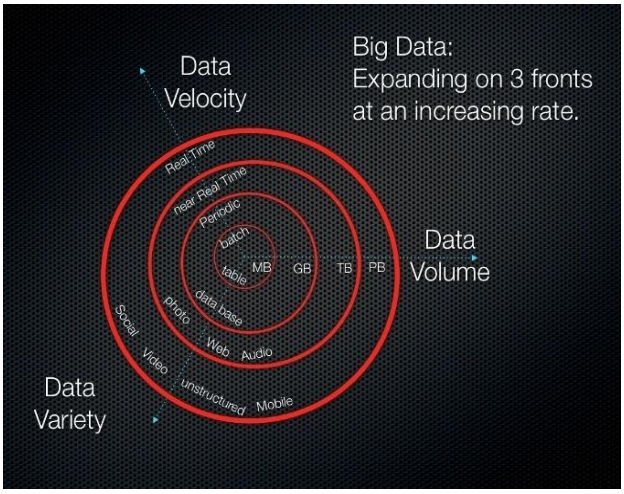

Figure 1:- Three V's of Big Data [1]

Some other characteristics are:-

1.4 *Variability* (Data in Highlight):-This factor refers to the inconsistencies in data which can create problem for Data Analysts in analysing the data properly. Thus it can hampers the process of data handling and managing effectively.

1.5 *Veracity* (Data in Doubt):-The data which is being captured have some quality and that quality of data can vary greatly. It depicts the trustworthiness of captured data. Veracity can affect the accuracy of analysis.

1.6 *Complexity*:-Data management becomes a difficult process when data is captured from various sources. All this data is linked, correlated and connected for getting the information that is meant to be conveyed by all these sources of data. When this scenario occurs the complexity is increased.

II. BIG DATA ISSUES AND CHALLENGES [6]

Big Data has some problems related to it which arises because of the increased demands of processing, storage and security of Big Data. These challenges cannot be handled by traditional and modern hardware separately. Some challenges are listed below that are addressed by Big Data:-

2.1 *Storage*:-Nowadays data is not just produced by human beings. Devices also produce data. Traditional and current disc technologies don't have sufficient capacity to store all the data produced together. Additionally, accessing all the data by users at single time can block communication networks. One solution of this problem is to remove redundant data for saving storage space. Detection of this unnecessary data from raw datasets and eliminating this redundant data is a key requirement to reduce storage overhead.

2.2 *Data Management*:-Data is generated at various sites which are geographically apart. This data is created, manipulated and managed by various organizations. This data can be present in various formats at different sites and this is also possible that needed information regarding this data may or may not be present in source metadata. The issues of accessing the data, its metadata, its references and integration of data are needed by the manager to manage the data at satisfactorily level. Management of data is difficult because of the high volume and variety of data available. It becomes difficult to validate the real time data because of high velocity of data.

2.3 *Data Representation*:-As data sets are being collected from various different sources of data. They can vary in formats and structure as well. To represent the data in a effective manner the integration of these datasets is required. The datasets are being integrated, stored in a uniform structure so that it can represent a valid valuable information. This representation can be same or different at different sites according to the needs.

2.4 *Risk Detection*:-Risks are analogous to contravention and violation of confidential data. Private data has to be protected from attackers. Securing the confidential data in a healthcare and financial department is a basic requirement. Information leaks and infringement can cause a great loss to business, company if its secrets and private data is disclosed to the outside world. An organization can suffer from financial loss due to this leakage. So, risk detection must be a proactive process. This means the consequences of breaches and infringement must be detected before occurrence. Risks must be detected as well as prevented from leakage.

2.5 *Security and Privacy*:-Today as volume of data is increasing at a fast rate. So, cloud technologies came into existence because of great need of processing and storage when demanded. This discloses the private data to outside and thus creating problems in the confidentiality of data. Hence, there is a need of securing this data from attackers and provide privacy to it. In this paper we concentrates on problems of security bothered by Big Data and will try to traverse directions of research.

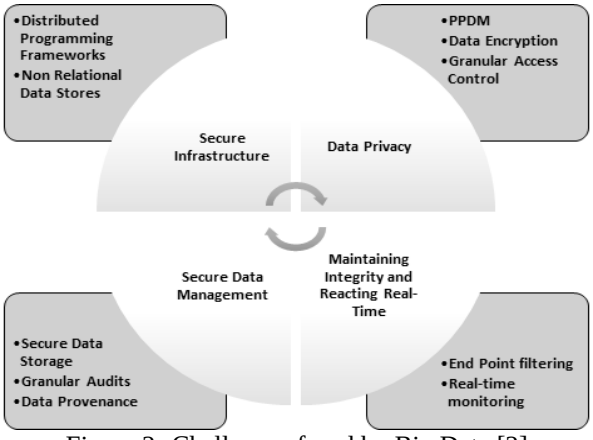

Figure 2:-Challenges faced by Big Data [2]

III. LITERATURE REVIEW

Occurrence of some kind of threat depicts the loopholes in privacy and security. As essence of privacy and security threat is dynamic in nature which makes a challenging task for providing security. Dynamic data is the most important part in Big Data. Thus, providing privacy and security during data transmission or while storing the data is the basic requirement of Big Data. Generally a data is called secured if CIA (Confidentiality, Integrity, and Availability) is achieved [7]. Ensuring privacy means hiding Personally Identifiable Information (PII). Hence, main privacy and security issues are confidentiality, integrity, availability, monitoring and auditing, key management and data privacy. While transmission data can be lost or changed by intruder which can be prevented by using some cryptographic methods [12].

3.1 Confidentiality

Confidentiality means preventing data from any authorized access. There are different techniques available to achieve confidentiality from which most popular are cryptographic techniques. The below said are the ways which can be used to sustain it in normal security mechanisms [8].

- First, data is encrypted while transmission but stored in the form of plain text.
- Access is proved to authenticated users only.
- Data is stored in encrypted form and decrypted while using.

Confidentiality is achieved by using AAA (Authentication, Authorization and Access control) [8]. Authentication means identity has been provided to the user [15]. In authorization the resources are being provided to authenticated user. Granting permissions to authenticated users for authorized use implies access control [10]. Many encryption methods for protection of confidential information using hardware are discussed in [7]. Seamless data's efficient authentication mechanism is proposed in [9].

3.2 Integrity

Integrity of data refers to the protection against alteration of data in any unauthorized manner by unauthorized user, Intruders, User errors, hardware errors or software errors are the key points for data integrity issues [8]. Someone well known attacks related to data integrity are man in the middle attack, data diddling attacks, salami attacks, trust relationship attacks [18]. Decision Making is very important in Big Data while trustworthiness has a great impact on analysis of decisions, actions and predictions [11]. Trustworthiness becomes difficult if there is no protection from malicious users [3].Data leakage or theft affects the data integrity and if data is copied it can create confusion [13]. Data Loss Prevention (DLP) approach can be used for providing protection from information loss. Xu et el. [7] has explored many approaches like data integrity protection by using digital signature, integrity protection of queries, integrity protection for storing information and by using hardware. Puthal et el. [15] provided security verification for real time streaming data. End to end security is provided by using dynamic key length.

3.3 Availability

This means data should be present whenever it is asked by authorized user. For providing data availability the High Availability (HA) systems are used [8]. While designing HA systems backup servers, alternative or duplicated transmission links have been used. However due to existence of cloud computing problem of data availability is decreased but still DoS attack, DDoS attack or SYN flood attacks can be the breaches for data availability [18].

3.4 Monitoring and Reviewing

Reviewing and Monitoring is needed for achieving security to identify abnormal activity in Big Data security system. Malicious and non-malicious data's monitoring behaviour is also necessary. As data is present in high amount so blocking data at server side or validating data at client side for reviewing is not the valid solution for Big Data. Reviewing is complicated because of dynamic nature of Big Data. While fine attempt for auditing of dynamic storage of information is discussed [16]. This reviewing approach overcomes the issue of authentication.

3.5 Key Exchange

In Big Data, Key exchange between users is a complicated problem. Many techniques for key exchange like sharing key secretly, approach based on server side and cryptography using signature exists [14]. Users are not required to keep key by themselves but they are supposed to share confidential data between various servers in Ramp secret sharing scheme (RSSS) [13]. Few secrets are used for generating key again.

3.6 *Data Privacy*

Data Privacy ensures that Personal information of any user should not be accessed by anyone without the permission of data owner. Ensuring Privacy means hiding Personally Identifiable Information (PII).If due to some reason the PII is shared than use of it must be restricted to that extent which is defined by the owner. A Generic framework PriGen for privacy protection in healthcare is explained in [17]. PriGen works on the data division method which checks and distinguishes sensitive data. Privacy can be protected by using schemes like k-anonymity, t-closeness, I-diversity [17]. Apart from this privacy protection by differential privacy and how to achieve privacy protection on client side is also discussed.

Table 1: Comparison of Security Issues

1. Issues	[3]	[11]	[7]	[8]	[9]	[10]	[12]	[13]	[15]	[16]	[17]
2. Confidentiality	-	-	Achieved	Achieved	-	-	-	-	-	-	-
3. Authentication	-	-	-	Achieved	Achieved	-	-	-	Achieved	Achieved	-
4. Authorization	-	-	-	Achieved	-	-	-	-	-	-	-
5. Access Control	-	-	-	Achieved	-	Achieved	-	-	-	-	-
6. Integrity	-	Achieved	Achieved	Achieved	-	-	Achieved	-	Achieved	-	-
7. Data provenance	-	Achieved	-	-	-	-	-	-	-	-	-
8. Data Trustworthiness	Achieved	-	-	-	-	-	-	-	-	-	-

9. Data loss	-	-	-	-	-	-	Achieved	-	-	-	-
10. Data redundancy	-	-	-	-	-	-	-	Achieved	-	-	-
11. Availability	-	-	-	Achieved	-	-	-	-	-	-	-
12. Monitoring and Reviewing	-	-	-	-	-	-	-	-	-	Achieved	-
13. Key exchange	-	-	-	-	-	-	-	Achieved	-	-	-
14. Privacy	-	-	Achieved	Achieved	Achieved	Achieved	-	-	-	-	Achieved

IV. CONCLUSION

As Big Data technology is enhancing day by day which means data is generate at a high rate. This data could be sensitive which needs to be shielded from the outside world. Challenges of Big Data like privacy and security are elaborated in this paper and further we explained confidentiality, integrity, availability, monitoring and reviewing, key exchange and data privacy in detail. Many good practices have done for resolving privacy and security issues. As Big Data is developing at a fast rate, it is necessary to pay attention towards privacy and security solutions of big data to make it more reliable to use.

REFERENCES

[1] https://www.datasciencecentral.com/forum/topics/the-3vs-that-define-big-data

[2] BigDataWorkingGroup, "Expanded Top Ten Big Data Security and Privacy Challenges," 2013. [Online].Available:https://downloads.cloudsecurityalliance.org/initiatives/bdwg/Expanded_Top_Ten_Big_Data_Security_and_Privacy_Challenges.pdf.

[3] Bertino, E. (2015, June). Big data-security and privacy. In *Big Data (BigData Congress), 2015 IEEE International Congress on* (pp. 757-761). IEEE.

[4] Katal, A., Wazid, M., &Goudar, R. H. (2013, August). Big data: issues, challenges, tools and good practices. In *Contemporary Computing (IC3), 2013 Sixth International Conference on* (pp. 404-409). IEEE.

[5] Praveena, M. A., &Bharathi, B. (2017, February). A survey paper on big data analytics. In *Information Communication and Embedded Systems (ICICES), 2017 International Conference on* (pp. 1-9). IEEE.

[6] Chandra, S., Ray, S., &Goswami, R. T. (2017, January). Big Data Security: Survey on Frameworks and Algorithms. In *Advance Computing Conference (IACC), 2017 IEEE 7th International* (pp. 48-54). IEEE.

[7] Xu, L., & Shi, W. (2016). Security Theories and Practices for Big Data. In *Big Data Concepts, Theories, and Applications*(pp. 157-192). Springer, Cham.

[8] Sudarsan, S. D., Jetley, R. P., &Ramaswamy, S. (2015). Security and Privacy of Big Data. In *Big Data* (pp. 121-136). Springer, New Delhi.

[9] Jeong, Y. S., & Shin, S. S. (2016). An efficient authentication scheme to protect user privacy in seamless big data services. *Wireless Personal Communications*, *86*(1), 7-19.

[10] Yang, K., Han, Q., Li, H., Zheng, K., Su, Z., &Shen, X. (2017). An efficient and fine-grained big data access control scheme with privacy-preserving policy. *IEEE Internet of Things Journal*, *4*(2), 563-571.

[11] Azmi, Z. (2015). Opportunities and Security Challenges of Big Data. In *Current and Emerging Trends in Cyber Operations*(pp. 181-197). Palgrave Macmillan, London.

[12] Gao, Y., Fu, X., Luo, B., Du, X., &Guizani, M. (2015, December). Haddle: a framework for investigating data leakage attacks in Hadoop. In *Global Communications Conference (GLOBECOM), 2015 IEEE* (pp. 1-6). IEEE.

[13] "SANS Institute InfoSec Reading Room", Sans.org, 2017. [Online].Available: https://www.sans.org/reading room/whitepapers/dlp/dataloss-prevention-32883.

[14] Jeong, Y. S., & Shin, S. S. (2016). An efficient authentication scheme to protect user privacy in seamless big data services. *Wireless Personal Communications*, *86*(1), 7-19.

[15] Puthal, D., Nepal, S., Ranjan, R., & Chen, J. (2015, November). A dynamic key length based approach for real-time security verification of big sensing data stream. In *International Conference on Web Information Systems Engineering* (pp. 93-108). Springer, Cham.

[16] Liu, C., Ranjan, R., Yang, C., Zhang, X., Wang, L., & Chen, J. (2015). MuR-DPA: Top-down levelled multi-replica merkle hash tree based secure public auditing for dynamic big data storage on cloud. *IEEE Transactions on Computers*, *64*(9), 2609-2622.

[17] Rahman, F., Ahamed, S. I., Yang, J. J., & Wang, Q. (2013, June). PriGen: A Generic Framework to Preserve Privacy of Healthcare Data in the Cloud. In *International Conference on Smart Homes and Health Telematics* (pp. 77-85). Springer, Berlin, Heidelberg.

[18] "Types of Network Attacks against Confidentiality, Integrity and Avilability", Omnisecu.com, 2017. [Online].Available:http://www.omnisecu.com/ccna-security/types-of-network attacks.php.

Authors Profile

Dinesh Singh Assistant Professor at Department of Computer Science and Engineering in Deenbandhu Chotu Ram university of Science and Technology,Murthal,India since 2006. He has total teaching experience of 13 years . His main research area is Signal Processing and Pattern Recognition.

Bharti Duhan pursued Bachelors of Technology from Guru Jambheshwar University,Hisar,India in year 2016 and currently pursuing Masters of Technology in Department of Computer Science from Deenbandhu Chotu Ram university of Science and Technology,Murthal,India. Her main research work focuses on Issues of Security In Big Data.

