Sundus *et al.LGU (IJECI)* 2020

LGU (IJECI)
ISSN: 2522-3429 (Print)
ISSN: 2616-6003 (Online)

LGU International Journal for
Electronic Crime Investigation

Research Article

Vol. 4 issue 4 Year 2020

Overview of Security Measures in Big Data

Sundus Munir[1] Afrozah Nadeem[2], Syeda Binish Zahra[3], Sadia kousar[4]
sundusm1@gmail.com afrozah@gmail.com,,
binishzahra@gmail.com[3], sadiakousar@gmail.com[4]

University of Engineering and Technology[2]
Lahore College for Women University[1,4]
National College of Business Administration & Economics (NCBAE)[3]

Abstract:

Big data refers to the huge volume of data that is being produced by different organizations after every second. It is hard to handle such volume of data, therefore this data is combined and managed through the servers of organizations using complex algorithms. Data is currently most important assets of any organization in every field with which many operations can be performed. Big Data exerts different properties those are volume, velocity, veracity and variety. Now there are different threats regarding the security of big data. Organizations are using different methods to secure big data. Big data security measures include architecture security, infrastructure security and data privacy. In this paper multiple security concerns are discussed and how they can help organizations to secure their data.

Keyword: Big Data, Cloud Computing, Hadoop, Encryption.

1. Introduction

Big Data is the term that describes the large volume of the data that is collected through different sources. Big data help the organizations to grow in different sectors. Big data is very complex as compared to traditional software's that is linked to data processing. Big data processing requires the system with more statistical power. Big data is not only about collection of data but also there are different challenges like how to capture the data, its storage, analysis, transfer and updation.

Big Data analytics tools and techniques are rising in demand due to the use of Big Data in businesses. Organizations can find new opportunities and gain new insights to run their business efficiently. These tools help in providing meaningful information for making better business decisions.

The companies can improve their strategies by keeping in mind the customer focus. Big data analytics efficiently helps operations to become more effective. This helps in improving the profits of the company. Financial data scientists use big data to predict

which stocks will succeed and when future crashes are likely to occur. Banks also see big data as a way to increase their revenue.

Big data grows rapidly because numerous of devices are interacting with system. Every interaction generates series of data bytes which includes transaction, reports, logs and documents. Since 1980 the world's technological per-capita [1] capacity to store data has doubled every forty months and IBM added a update that since 2012 the data generation is increased 2.5 extra bytes.

Big data contains sensitive data of organizations and individuals. For example, companies that provides cellular services collects data of numerous Calls, SMS and internet data which is directly linked to the privacy of the consumers. Company use this data to store, analyse, update, search, transfer through complex machines. This arises a alarming thought that if some third party breach into big data and collects sensitive information it might cause a big problem to the privacy of customers. Big data security is very crucial that the US National government financed $200 million to Big data research [2]. This portray that big data is expanding vastly and exerting its impact on daily life.

Big data is making a huge difference in many fields like healthcare [27]. Insurers and providers are working on combining data from different sources such as claims, X-rays, doctor's notes, and prescriptions. Many believe that if healthcare data better integrated it could provide better care at a lower cost.

2. What is Big Data?

Big data is used when there is enormous volume of data regardless it is structured or unstructured. The traditional systems and databases cannot handle this enormous volume of data. It requires a massively parallel software running on hundreds of server systems. Big data can be characterized and understood with the help of 4V's which not only exhibit the complexity of big data but also its speed [3]. The 4V's are

Volume
Variety
Velocity
Veracity

2.1 Volume:

The term volume in big data is used to show the amount of data that is generated by some company or organization. The quantity of data not only refers to the value of data but also potential of the information derived from the data. The collection of massive data in structured and unstructured form required connection of servers with the dynamic storage capabilities. Several Companies like Amazon [4] is providing storage services which may vary from user or company requirement.

2.2 Variety:

In big data different software's and systems are connected and they generate different data in various formats. For instance, a media company generate several copies of audio, video files while a software house generate data of different formats like logs, sheets, software versions and some unstructured data which is not being explored using traditional software's. This shows that how much variety of data is processed in big data.

2.3 Velocity:

To understand what velocity is in big data there are two situations. One is the speed of generation of data with in a system and second is speed of data storage. Like a telephone

company generate hundreds of thousands of data like logs and files in seconds. Hence, it is about to meet the demands of speed. The term velocity also refers to the speed of the data processing that is stored in the systems.

2.4 Veracity:

The term veracity refers to the trustworthiness of big data. The uncertainties that are found in data because of incompleteness, deception and ambiguities reduce the quality of data. The quality of data that is stored depends on the analysis of data which is directly linked to the Veracity of data.

3. Big Data Analysis

In big data analytics mostly the process of data mining and analysing occurs. This process can help to achieve business knowledge and operational skills to unprecedented scale. Big data analysis helps to identify the trends in the data collected by business. Big data analysis helps in examining chunks of data which results in finding the un-known correlations, improved customer services, better business strategies, marketing patterns and efficient customer support. It includes different processes that are predictive analysis, text analysis. Here are some of the advancements that are occurred due to the field of big data.

- Cloud Computing [5] and data centres acquired flexible storage and computing resources to manage and apply operations on data collected from the business.

- Different frameworks are designed to manage and store large quantity of data. Best example of the framework is Hadoop which helped multiple users to get advantage of cloud computing [6].

- Storage cost has been reduced since last few years that a user can acquire hundreds of giga bytes from computing and data storage purposes in few bucks.

- Big data is very helpful in machine learning. By following the results and patterns of big data new algorithms for machine learning has been designed.

- Big data has created many job opportunities like Big Data Analyst, Big Data Engineer, Business Intelligence Consultants, Solution Architect, etc [7].

4. Stages of Big Data

Here are few stages of big data that are followed in the process of making the data useful for any particular goal. After following these stages organization can gather valuable piece of information from data.

4.1 Understanding of Business:

Goals are necessary to understand that what is best for achieving business point of view. In the business how data can be beneficial and what measures and parameters should be defined before processing the data. The problem statement is target and then Decision Model is utilized for the goal achieving.

4.2 Understanding of Data:

For this purpose, data is collected for understanding so that information derived from data is accumulated. Data understanding also refers to the mining of data. After data is organized it can be used in achieving the goals.

4.3 Preparation of Data:

Data is prepared in the forms of tables and databases so that the quality data is acquired in the fastest way. Data is manipulated in the structured form like table to make it suitable

for next processes.

4.4 Data Modelling:

After all these steps a data mining model is selected. In data modelling it is easier to understand that how data flows to complete the system with the help of symbols and diagrams. Model is consisted of a scheme that fulfils the requirements of parameters in the systems.

4.5 Data Evaluation:

In data evaluation every chunk of data is reviewed with the deep analysis and logical reasoning. A model is selected for data processing and after that a survey is organized to make it clear that the model will successfully accomplishes the goals of the business.

4.6 Deployment:

Last step is the deployment of the model to finish the venture. During deployment of data Hadoop Tool is used to handle the changes of big data. In deployment data is being processed and analysed for the results [8].

5. Data Visualization in Big Data

In modern industry IOT (Internet of Things) refers to adopt new policies and methods. This not only result in generation of variety of skills and also reducing global competition stress regarding data handling. The total volume of data after internet of things has come increased to Thousand times more than the last decade. This revolution after inclusion of data in the industries and digital ecosystem is called Fourth Industrial Revolution (add reference). More devices that are connecting to the system are tools, plants, machines, auto mobiles, robots and they are producing data at very high rate. This data is being utilized by the big companies enabling to unlock the untapped possibilities in every field of life. The idea to produce fault free machines can be pursued in the light of big data because it will help to acquire the best performance levels. Data Visualization is big challenge in big data. There are different approaches that are being used for the data visualization and also multiple tools that are as follows [9]

- Plotly
- DataHero
- Tableau
- ZingChart
- Chart.js
- Google Analytics
- Dgraphs

These tools are being used for data visualization. For real time understanding and reasoning of big data, we work with supply chains to produce results i.e., manufacturing intelligence. [10]

6. Big Data Security

Within the new era of technological revolution there are also some issues that arise and some are becoming strong challenge for the technology. The same happened with the revolution of internet of things and big data. These issues are deeply linked to the volume and privacy of data. Traditional security measures, tools and management are not proper solution of these problems. Without building proper solutions of the problems arising in big data it will not achieve the required level of trust. According to group of people working on Big data in Cloud Security [11] Alliance stated that there are majorly four different aspects of Big data security that are as follows

- Infrastructure Security
- Data Management
- Data Privacy

- Integrity and Reactive Security

In the International Organization of Standardization these four topics have been used in the area of Big Data Security [12].

6.1 Infrastructure Security

There are technologies and frameworks that are being used in the infrastructure of big data to secure it. The discussion of these frameworks and technologies are must for big data security and especially for those which are purely based on Hadoop Technology. Most of the big data systems are based on Hadoop. Further following are some important points which discussing the infrastructure security [13].

- Security for Hadoop
- Architecture Security

6.1.1 Security for Hadoop:

When we deal with the infrastructure security Hadoop is necessary. Hadoop platform uses map reduce programming model to process the data.[14] Researchers have proposed some models for the security of Hadoop in which includes the use

of new schema and creation of encryption keys and schemes. Hadoop ecosystem has a framework that is known with the name of Knox whose solemnly function is to manage security implementations across the Hadoop cluster. In Hadoop File Distributed system (HDFS) the encryption process works in such a way that there are zones defined and each zone is created by different directories. [15] Each zone is encrypted by some key. That unique key is called as DEK (data encryption Key) only client can decrypt the protected data and can use it. Hadoop has its unique feature of storing all keys in a server. So, this encryption key feature helps Hadoop to authenticate clients and data [16]. There is a scheme for data confidentiality that is called as Trusted Scheme for Hadoop Cluster (TSHC). This TSHC creates an architecture for Hadoop system which is improve the infrastructural security for big data [17].

6.1.2 Architecture Security:

Rather than depending on a single architecture the security of the environment. HDFS with the combination of network coding and multi node reading there is less chances of vulnerability. Here is also a suggestion that the data centres should be built near to the data so that the data communication would face less risk. Sensitive fields should be encrypted with the algorithms and keys so that risk of manipulation and exposure of data to unknown entity is reduced. When data is encrypted, important point is key management. Such a mechanism should be implemented that keys will be generated on base of need.[18]

6.2 Data Management

Once all the data that system generates is collected in servers of big data. Here a new point arise how it should be managed. In order words what we need to do with data and if it is viewed in the sense of security how to securely manage it. These are some important points which fall under data managements. What are the security measures around the servers where data is collected? [19]. What are the security parameters when all this data is being stored in the servers? The main point is to protect the user's privacy and for that some parameters and levels should be set.

A very serious problem around big data that is how this data can be used. Different companies make different policies regarding data use. They can make it open for their clients if they agree to use this data purposely with restrictions. Recently an incident occurred

"Facebook–Cambridge Analytica Data Scandal" [20] in which data of public profiles and their interests have been used without their consent. That data has been used indirectly to manipulate the election results. This clearly shows that how data is valuable and if it falls in wrong hands then it can be fatal for not only an organization but also threat to entire nation. After this incident several governments apply new rules and govern some strict laws to protect the data because it is their privacy which is at stake mostly. Organizations have been working to create some new rules that will reduce this risk of mismanagement of data.

The key thing in the data management is sharing and without sharing there is no purpose of the entire system. Large amount of data is collected and shared on the servers and sometimes different companies with big data collaborates to launch a new service or product to handle risk and threats. so the risks and threats to data should be ended.

6.3 Data Privacy

Data Privacy refers to which and how data (stored in system) is shared with the third parties. Privacy is the most important measure linked to the people could be at stake [21]. This matter is also important to the organizations and companies who deal with big data and using it for their advantage. There is a thought emerging in researchers that there must be rules and laws to limit the use of the data. Companies should be allowed to use the data for their benefits but also secure the privacy of collected data. The privacy requirements should be specified for collecting, storing and processing the big data.[22]

There are different ways to secure the data and more ways are being introduced. One of them is cryptography and it is the frequently used method. The most famous cryptographic algorithms are Advanced Encryption Standard (AES) and RSA Algorithm. Along with these algorithms some other methods which would be implemented like firewalls, transport layer security and they act as virtual barrier in the access of data. The software's which are developed for the purpose of surveillance and tracking are very complex to implement on very large data. Along with the risk of exploitation of these software's, they require a staff with good technical skills and huge implementation cost. The very next and important point is access control for the data privacy in which the users who doesn't have a role towards the data should be blocked or restricted. Some authors refer to some frameworks in order to manage the access control features and some gave importance to the map reduce process.[23] The approach to data has to be confidential and for that purpose new techniques are proposed. CMD (Computing on Masked data) this scheme allows computations to be performed on masked data which will not only improves the integrity but also data confidentiality.

There are different ways to secure the privacy of data by making it anonymous. Such kind of mechanism should be used that either hide the data or remove the sensitive information automatically. There are two schemes which are used to make the data anonymous. First one is Top -Down Specialisation, it is natural and efficient for handling both categorical and continuous attributes. This top-down specialization data usually removes redundant structures for classification to make the data anonymous [24]. Second one is Bottom-up generalisation, this approach incorporates partially the requirement of a targeted data mining task into the process of masking data so that essential structure is preserved in the masked data. Once the data is masked,

standard data mining techniques can be applied without modification [25].

People are also using social media network and its popularity is increasing day by day. Different platforms of social media are introduced with different features. Organizations who run social media are well aware that how much data is collected and processed in their servers. Privacy is the most important aspect in the social media networks. [26]

6.4 Integrity and Reactive Security

One of the bases on which Big Data is upheld is the ability to get streams of information from various sources like in distinct format from different origins: either structural or un-structural data. This builds the significance of checking that the data's integrity is upright so it may be utilized appropriately. It can also be used to monitor security so as to detect whether system is attacked or not.

Integrity is very important it defines the trustworthiness, accuracy and consistency of data. During its life cycle it protects data from unauthorised modifications. It is one of the most important concepts in dimension of security. In Big Data environment, achieving integrity is very critical when attempting to manage different problems.

Conclusion

This paper concludes the importance of big data. How data is being organized, collected, processed and what are the security threats. The next era of technology is mostly based on how organizations, Government and private institutes are using data for different goals and purposes. Therefore, multiple security mechanisms are suggested for companies. Each organization has set different steps and criteria to process the data into useful information. Security of big data involves the security of infrastructure, data management and privacy of data. In infrastructure security the Hadoop tool security is key point because it processes and communicate with the clusters of data. Architecture security is also important in every aspect in which we see what schemes are used, what are the locations, where physical systems are located and where data is being processed. Best way to secure data is to build data centres where data is being collected. Data Privacy is very important in big data security as previously explained the scandal of Facebook it shows how data can be used for manipulation of ideas and thoughts of people. Millions of people are somehow connected to systems and they are sharing information which is processed in some online systems. Data management is also important point in big data because if data is not managed properly with proper parameters, then it would give the results which are not required to achieve the goal for an organization. Organizations need to focus on the big data security if they really want to grow the company and build trust in the market.

References

[1] IBM What is big data? – Bringing big data to the enterprise. www.ibm.com. Retrieved 26 August 2013

[2] K.Valli Madhavi, Dr.Y.Venkateswarlu ,Varsha Sharma ,Big Data Analytics for Security, 2018

[3] P. Kamaksh, Survey on Big Data and Related Privacy Issues, Voulme 03,Issue 12, pp.68-70,Dec 2014

[4] Cloud Storage Services–Amazon Web Services (AWS) – aws.ama-

zon.com/products/storage. Retrieved August 2014.

[5] Big Data Security – The Big Challenge Minit Arora, Dr Himanshu Bahuguna, 2016.

[6] Devaraj Das, Owen O'Malley, Sanjay Radia and Kan Zhang ―Adding Security to Apache Hadoop.

[7] Why is Big Data Analytics So Important? www.whizlabs.com/blog/big-data-analytics-importance - Retrieved 19 March 2018 .

[8] SURVEY OF BIG DATA SECURITY Snehalata Funde, Computer Engineering, BSCOER, Narhe, India.

[9] Big Data Visualization Tools Everyone in the Industry Should Be Using - Promptcloud.com, Reterived 24 Feburary 2016.

[10] A, Katal, Wazid M, and Goudar R.H. "Big data: Issues, challenges, tools and Good practices."., pp. 404 – 409, 8-10 Aug. 2013.

[11] W. Hao, "Secure Sensitive Data Sharing on a Big Data Platform", Tsinghua Science and Technology, vol. 17, no. 1, (2015), pp. 72-80.

[12] Wang, H.; Jiang, X.; Kambourakis, G. Special issue on Security, Privacy and Trust in network-based Big Data.Inf. Sci. Int. J. 2015

[13] A Novel Framework for Big Data Security Infrastructure, Manpreet Kaur, Saravajanik College of Engineering and Technology.

[14] Cugoala, G. & Margara, A. (2012). Processing Flows of Information: From Data Stream to Complex Event Processing. ACM Computing Surveys 44, no. 3:15.

[15] Priya P. Sharma, Chandrakant P. Navdeti, (2014), " Securing Big Data Hadoop: A review of Security Issues, Threats and Solution", IJCSIT, 5(2), pp2126-2131.

[16] Cohen, J.C.; Acharya, S. Towards a trusted HDFS storage platform: Mitigating threats to Hadoop infrastructures.

[17] Quan, Z.; Xiao, D.;Wu, D.; Tang, C.; Rong, C. TSHC: Trusted Scheme for Hadoop Cluster. In Proceedings of International Conference on Emerging Intelligent Data &Web Technologies (EIDWT), Xi'an, China, 9–11 September 2013.

[18] Frank, J.B.; Feltus, A. The Widening Gulf between Genomics Data Generation and Consumption: A Practical Guide to Big Data Transfer Technology. Bioinf. Biol. Insights 2015, 9 (Suppl. 1), 9–19.

[19] Wang, Y.; Wei, J.; Srivatsa, M.; Duan, Y.; Du, W. IntegrityMR: Integrity assurance framework for big data analytics and management applications. In Proceedings of the 2013 IEEE International Conference on Big Data, Silicon Valley, CA, pp. 33–40.

[20] 50 million Facebook profiles harvested for Cambridge Analytica in major data breach – www.theguardian.com, Retrieved 18 March 2018.

[21] What is Data Management? – NGDATA https://www.ngdata.com/what-is-data-management/ Reterived 31 March 2016.

[22] Xu, L.; Jiang, C.; Chen, Y.; Ren, Y.; Liu, K.J.R. Privacy or Utility in Data Collection? A Contract Theoretic Approach. IEEE J. Sel. Top. Signal Proc. 2015, 9, 1256–1269.

[23] Sultan Aldossary, William Allen, "Data Security, Privacy, Availability and Integrity in Cloud Computing: Issues and Current Solutions", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 4, 2016.

[24] Top-Down Specialization for Information and Privacy Preservation, Benjamin C. M. Fung, Ke Wang, Philip S. Yu.

[25] Bottom-up generalization: a data mining solution to privacy protection, Ke Wang ; P.S. Yu ; S. Chakraborty.

[26] Sundus Munir "Social Media and its Impact on Privacy", International Journal of crime Electronic Investigation (IJECI), (2018).

[27] Afrozah Nadeem, Sundus Munir, Syeda Binish Zahra, Sadia Kousar, "Challenges and Opportunities of Big data in health care" , International Journal of crime Electronic Investigation (IJECI), http://ijeci.lgu.edu.pk/index.php/i-jeci/issue/view/11/11