International Journal of Engineering & Technology, 7 (1.3) (2018) 100-103

International Journal of Engineering & Technology

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

Research paper

Big data life cycle: security issues, challenges, threat and security model

Bhima Sankaram Alladi [1] *, Dr. Srinivas Prasad [2]

[1]*Research Scholar, Dept. of CSE, Koneru Lakshmaiah Education Foundation, Guntur, Andhra Pradesh, India*
[2] *Professor, Dept. of CSE, Koneru Lakshmaiah Education Foundation, Guntur, Andhra Pradesh, India*
**Corresponding author E-mail: dhana1206@gmail.com*

Abstract

Today the technologies of big data are completely bringing a vast change in the entire conventional technology discipline and it's successfully applying the required latest security design methods to state the upcoming security provocations. Big Data Architecture is a "Data" centric architecture in which security can be included in all the levels. Data is collected from different sources and Data generation is done, the next step it undergoes is Data Processing, the next step is Data storage and the last step is Data analysis. At all the levels Data plays a vital role. It aims to give basic investigation regarding most of the security risks and Big Data provocation and bought out new provocations, complication to the conventional protective domains and also for conventional trends. This deals with the definition of big data and the characteristics that effect most of the data preservation, such as 3V's, dynamicity. It analyses the original changes and new challenges to Data security. It also provides pitch for real time practice of security infrastructure peripherals which allows extend trusted non-local virtualized processing environment. This research focus on all levels of Big Data where and when the security services and techniques can be included to acquire accurate results.

Keywords: *Big Data; Challenges; Issues; Privacy; Security Life Cycle.*

1. Introduction

From past few years there is a huge demand in technologies. The development was associated with gigantic measure of bit organized data. The terabytes and terabytes of information has been expanded to Peta bytes and Zeta bytes. This information expands day by day growth without breaking point. Currently the entire world is completely converting from human made to automation, (i.e.,) from a small retail store to the central government everything is converted to online. Each cooperation and correspondence being automated has become a huge source of information hub. This transformation has brought huge demand for the approaches of Big Data technologies [1].

Data incorporates an extensive variety of information like trade exchange, stock exchange market, association files, healing facility logs, scholar information, private data, movable data, geographical information, CC chronicles, etc. This field produce vast amount of data at a rapid growth. This information creates conflicting, shifted in environment implies they are not organized [1].

As with the rapidly growing data, the issues identified with Big Data are additionally expanding. A few issues of big data to be tended like analysis, storage and management of data, etc. This significant risk need to be managing the data in security and privacy of big data. This issue mainly raised due to the natural of big data features [2]. This worry fundamentally emerges because of the inborn idea of Big Data characteristics. There are demonstrated security existing components, we concentrate on utilizing current security systems to beat the impacts of big data features.

The rest of the paper is aligned as follows: In section 2, we represent the big data characteristics. In section 3, briefing the outline of Challenges and Issues in Big Data. In section 4, we represent the possessions of big data characteristics on security. In section 5, outlining the big data security Lifecycle and in section 6, mentioning about security threats and attacks in big data. Followed by conclusion and future Scope [2].

2. Big data characteristics

The 3-dimensions which describe data as 3 V's majorly later it has now been to 7 V's, used to define the data characteristics as follows [3].

i) Volume

The Volume defines the measure of data. The volume big data era has crossed terabytes and petabytes, getting to be plainly measure to analyze different big data sets. The huge volume of information is expanded step by step at a wide range.

ii) Variety

The characteristic variety explains the broad range of data formats encircled by big data. It consists of different natures of data like audio, video, images, text and raw data, arithmetical information, mail data, etc. It also comprises of information from distinct fields like stock exchange, health and life science, organization information, scholar information, jurisdiction information files, geographical information, astronomer's data etc.

iii) Velocity

The characteristic velocity defines the speed of generating data (i.e.,) the growth at which the information is being generated. Advanced technologies with high data rate which in turn produces results quickly. Day to day, the rate of internet users has been increased globally in turn effects the production of data at every second. Many other societies, organizations make use of extra dimensions to produce the data variability

iv) Variability

The characteristic variability gives non-consistency of data at any point of time. At any point of time the data production may be high or low. The unexpected behavior of the data is a big concern.

v) Veracity

The characteristic veracity shows the good measure of restored data in the environment of data. Globally produced data with rapid speed has different parameters. Users of big data may vary in different concerns.

vi) Volatility

The characteristic volatility defines how long the information is appropriate and for how much time period it has to be stored. Most of the times data is valid for certain time period and became invalid after stipulated period.

vii) Validity

The characteristic validity illustrates the accuracy of data. Storage of erroneous or in correct information yields to a stage of break. The characteristics presents numerous differences between basic systems and data sources like volume, variety, scalability, data storage, data security, normalization, data redundancy and data analysis, etc. The present study is on security and its solutions and also discuss about the services for basic systems.

3. Challenges and issues in big data

Collectively term big data refers to data that is very huge and complex that it exceeds the processing capacity of traditional information management systems and processing techniques for software's [4]. There are many risks and challenges related to data security and data privacy and reasons where explained in detailed. The following are the areas:

- Big data risks and challenges related to features of big data
- Big data management, man power and human resource risks and challenges
- Big data technical issues and challenges
- Big data storage and transport risks and challenges
- Big data transforming risks and challenges
- Big data privacy and security risks and challenges

Among these we focus on information security challenges and issues as we relate this paper more with security issues more.

3.1. Privacy, security risks and challenges

Frequently in information analysis, the public's personal information from a database or from social sites need to be collectively joined with exterior huge data sets.[5] Generally, it leads to enchanting insights in human's life of which they are obvious of. Normally it happens that a more sophisticated individual having best awareness about advantages of bigdata and predictive analysis.

3.2. Reasons for security issues and challenges in big data

The two major concerns of big data are security and privacy, as it grows by its volume day by day, each and every moment; these concerns are on the raise [6]. As far as security and privacy is concerned the major senses in big data, because data is broadly available in recent days. Data are collectively shared on a significant by various people like doctors, scientists, business people, government officers, ordinary people and so many others. On the other hand, the tools and technologies that have been designed till now to handle these large amounts of data is not capable to provide sufficient privacy and security to data.

The technologies require security and privacy features, because they require the necessary understanding and how to provide security to these large amounts of information and adequate practice need to be provided regarding the maintenance of security and privacy to the large scaled information [7]. The maintenance of information security and privacy regarding big data which requires adequate framework that make sure the deal with most recent methods to risks of data. The current trends have least ability to maintain security and privacy issues. So, they are regularly being infringe both by co-incidence and purposely. Thus re-assessing and enhancing available methods to put off information outflow which has to be done on permanent basis. Instead of exhaust on information security in order to secure data by the firms, More than 10% of the organizations funds, is been drained for protecting their companies data.

4. Impact of data aspects on security

To tally huge quantity of information in a diverse environment, we can pre-own the approach like Map Reduce. The function of the mapper is to read information and reducer is to produce result. Usually in various cases, the mapper could be unreliable and it may include unrealistic information. Detecting such unrealistic information for huge amount of data is itself a very big task [8].

Generally in cloud environment, the information records are secured in various levels of tiers. At the beginning stage, the data has been originated for auto-tiering to accumulate the bulky volume of data. The security challenges such as untrusted security services and no restriction on data security location are familiarized.

An IDS in big data environment is a provocation for the production of numerous notifications for huge information repeatedly continued by various false positives. Maintaining this large amount of false positives in these huge quantities of information is itself a big provocation [9].

Privacy preserving technique is always be a prioritized challenge. The huge quantity of information collected contains the personal information. An un-reliable individual will get admission to the complete quantity of information without any effort which in turn effects the giving up their individual information.

Conventional systems maintain records and transactional files. Physical securing of transactional files contains professionals touch on the information. The performance of data volume contains mess up activities for physical filing in huge information.

The major attacks in security for the conventional methods are considered and a framework is generated using the data. The procedure of framework generation, techniques and associated relations is vital to put into practice for best benefits. For above generated framework analysis, the quantity fabricated is an exhausting work.

Firms hire professionals for live project monitoring the security actions. Practice is probable for conventional methods as the information is relatively not up to the mark.

Encrypting huge amount of information is a long procedure and occasionally might affect performance issues. Conventional method design steganography to look after mostly pictorial information. The broad range of information comes along some rival while designing steganography.

Filtering and validating in Big Data is a challenge because of its large amount of data. The end user has the capability to create malicious information and it can be can be gathered and restored for future use [10].

In the phase of data generation, the information can be gathered from wide diversity of resources. Mostly, databases were intended to hold risks related analysis but when coming to issue of security it is been completely give up due to diversity.

5. Big data security lifecycle

Now we present lifecycle model for data security and the major elements of any big data framework. We enhance our model from Xu et al. 2014. This paper deal with vast information from end user point of view, where they are 4 different categories of clients' role in this environment they are: provider, collector, miner, and decision makers for information. On the other hand, this design states the levels in data life cycle. Big Data security lifecycle

model subsist of the following 4 sections in the information approach which consists of data phases like collection, storage, processing, analysis, and knowledge creation. The following fig 1: states major components in data lifecycle [11].

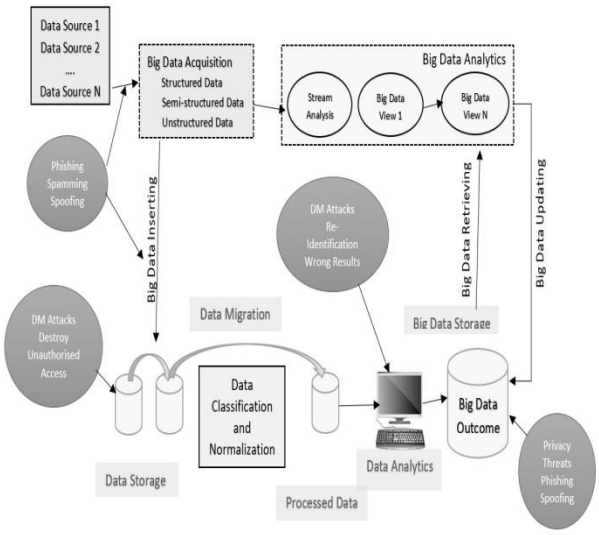

Fig. 1: Life Cycle Model for Big Data Security.

Data Collection phase

In this, the information from diverse sources generates various formats of information: structured (information with high degree of organization), semi-structured (information with moderate degree of organization), and unstructured (information with lower degree of organization). From a security point of view, securing big data technology should begin from the initial stage of this model. It is vital to collect information from reliable data sources and ensure that this is protected and secured. In reality, consider security procedures to track information from being leaking. Various metrics can be applied in data collection like limited access control and encode important data records.

Data Storage Phase

In this phase, the collected data from the data collected phase is secures and arranged for being applied in the upcoming stages i...e data analytics section. Usually the gathered information may include personal data, so it is necessary to have adequate safety measures while securing the information. In spite of giving assurance, the protection of the gathered information, of the secured metrics like data anonymization approach, data partitioning (vertical partitioning/ horizontal partitioning), and permutation are used.

Data Analytics Phase

Soon after the data collection is completed it is stored in secured storage format, then the data analysis is done to produce important information. The mining techniques like classification, clustering, and association rules are applied. It is critical to present secured processing environment. Actually, miners apply highly trusted mining techniques which can mine personal information. Hence, mining methods and its results must be sheltered in contrast to mining-based attacks and do definite that only respective employees can have access to this.

Knowledge Creation Phase

The preceding section contains innovative and honored data which is been applied by decision makers. The generated data is allowed as responsive data, especially in a competition environment.

6. Security threats and attacks in big data

Big information has start off with varied security threats and attacks. Massive information threats and attacks are broadly derived from the characteristics of huge information technology that have confidence information analytics techniques including data processing algorithms. In reality, attackers may also use data processing strategies and procedures to find responsive information and discharge it to community and so data violation happens. During this paper, we tend to reason threats and attacks of huge information in terms of the four phases of huge information lifecycle. The Table one explains that the threats and attacks at every part [12].

From the below table, we have a tendency to reason the threats and attacks supported the massive information security lifecycle phases. Every life cycle section has special characteristics and assigned numerous tasks, therefore every life cycle section is exposed to distinct threats and attacks. During this regard, the initial section information assortment is hospitable many attacks like phishing and spoofing attacks [13]. These forms of attacks are targeting public World Health Organization add assembly and providing information to huge information framework. A technique to boost security in information assortment section is to supply security awareness programs to information assortment workers and instruct them the way to change with security policies and procedures.

After knowledge assortment is completed, knowledge is hold on in knowledge storage devices; we've got to be attentive of some threats and attacks. During this regard, hackers World Health Organization get admission to info in storage devices might use data processing techniques to mine sensitive info and use it illegitimately. This type of attacks is referred to as data processing based mostly attacks. So as to subsume this type of attacks, we have a tendency to might divide the datasets vertically or horizontally to decrease the force of this attack and assume non-central info storage framework. On hand there are other threats associated with knowledge storage section like attacks on knowledge storage devices (ex. Stealing laborious disks) and unauthorized access attacks. During this regard, we will build a considerable guard and construct access management protocols [14].

Table 1: Security Threat Model

Phases	Threats and attacks	Description	Suggested defense
Data Collection	Phishing	These attacks are hacking data provider and collector to get an access to the data in the collection phase.	Security awareness program
	Spamming		
	Spoofing		
Data storage	Data mining based attacks	Targeted datasets to extract knowledge (Dev et al. 2012).	Divide datasets (vertically and horizontally) and non-central data storage framework.
	Attacks on data storage devices	Stealing hard disks or make images of them	Physical security measures noncentral data storage framework.
	Unauthorized data access	People access data illegally	Access control
Data analytics	Data mining based attacks	Using data mining methods to extract sensitive knowledge.	Divide datasets (vertically and horizontally) and use access control.
	Re-identification threat	Identification threats of personal information (Jensen 2013)	Core attribute encryption
	Wrong result threat	Using incorrect analysis process, which lead to incorrect results (Jensen 2013).	Follow correct analysis procedures and document, audit, and review the process
Knowledge creation	Privacy threats	Releasing the resulted knowledge (ex. Rival competitors)	Adopt encrypt the resulted knowledge and adopting access control strategy.
	Phishing and spoofing	Decision makers are targeted	Security awareness programs

In information analytics part, some threats and attacks could happen to discharge sensitive info or injury the knowledge method. data processing primarily based attacks could occur to seek out out and build public all the sensitive information or correlation techniques may well be wont to re-identify individual information that impact individual's information privacy. So as to seem once massive information framework from this sort of attacks, we have a tendency to could implement some defense procedures like dividing information sets into many elements (horizontally or vertically) and perform encoding to the core attributes (attributes with

high weight). On the opposite hand there's another threat during this part that's obtaining faulty results from the info analytics method. Therefore, it's essential to follow approved analytics method and document it for the longer-term use.
Finally, within the last part data creation we've got to think about the threats and attacks and the way to shield it. In reality, the created data from huge information method is taken into account responsive information that desires to not be discharged to the community and notably to rival corporations within the business context [15]. Some security attacks and privacy threats which will aim the choice manufacturers and people WHO have access to the ultimate outcome of the massive information method. Therefore, we want to style security policy and follow access management procedures furthermore developing security awareness programs to avoid and diminish the impact of any threat.

7. Conclusion

It is necessary to be attentive of security threats and attacks of data. Big data refers to the process of handling enormous bulk of data. With the ever-growth usage of big data, several provocations are constructed; particularly security provocations that mostly effect data privacy. This research focused on security threat life cycle method for large data and illustrates the security threats and attacks of large data in terms of the life cycle procedure. Data security life cycle model subsist of 4 sections namely collection, storage, analysis, and knowledge creation of data. Initially data collection section subsists of gathering data from diverse origin that is essential to gather data from reliable data sources. Secondly data storage phase, its main task is to collect the output of data collection phase and stored it by using secure data storage solutions. The 3rd section of our life cycle model is data processing, which in need to make sure how to keep information assurance during data processing. Mostly big data depends on data mining techniques to mine sensitive information. Hence, our future scope is to decrease the effect of the data mining-based attacks by mounting valuable security parameters like storage division and encode selected attributes of the data sets.

References

[1] Big data analytics: organizational factor matters impact technology acceptance, Journal of Big Data, Springer Open Access, June 2017.
[2] Big data privacy: A technological perspective and review, Journal of Big Data, Springer Open Access, November, 2016.
[3] An efficient strategy for the collection and storage of large volumes of data for computation, Journal of Big Data, Springer Open Access, October 2016.
[4] Big Data for Supporting Low-Carbon Road Transport Policies in Europe: Applications, Challenges and Opportunities, Big Data Research, Elsevier, December 2016.
[5] Reference Architecture and Classification of Technologies, Products and Services for Big Data Systems, Big Data Research, Elsevier, December 2015.
[6] Big Scholarly Data: A Survey, IEEE Transaction on Big Data, Vol 3, No.1, March 2017.
[7] Big Data for Cyber Security: Vulnerability Disclosure trends and Dependencies, IEEE Transaction on Big Data, Vol 3, No.1, October 2016.
[8] Methodologies for cross-domain data fusion, IEEE Transaction on Big Data, Vol 1, No.1, January 2015.
[9] Efficient big data processing in Hadoop MapReduce, Journal proceedings of the VLDB Endowment, ACM, Vol 5, Issue 12, August 2015.
[10] Z. Wu et al. "Towards building a scholarly big data platform: Challenges lessons and opportunities" pp. 117-126 2014
[11] Y.-R. Lin H. Tong J. Tang K. S. Candan "Guest editorial: Big scholar data discovery and collaboration" vol. 2 no. 1 pp. 1-2 Jan.-Mar. 2016.
[12] S. Kaisler F. Armour J. A. Espinosa W. Money "Big data: Issues and challenges moving forward". IEEE 46th Hawaii Int. Conf. Syst. Sci. pp. 995-1004 2013.
[13] S. Sagiroglu D. Sinanc "Big data: A review". IEEE Int. Conf. Collaboration Technol. System pp. 42-47 2013.
[14] J. Dean and S. Ghemawat. MapReduce: A Flexible Data Processing Tool. CACM, 53(1):72–77, 2010. https://doi.org/10.1145/1629175.1629198.
[15] S. Blanas et al. A Comparison of Join Algorithms for Log Processing in MapReduce. In SIGMOD, pages 975–986, 2010. https://doi.org/10.1145/1807167.1807273.