ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

Big Data: A Review on Security Perspective-Problems and Applications

Shawn Rahul D’souza[1], Nisha Jenifer Roche[2]

Under Graduate Student, Dept. of CS&E, St Joseph Engineering College, Vamanjoor, Mangaluru, India[1]

Assistant Professor, Dept. of CS&E, St Joseph Engineering College, Vamanjoor, Mangaluru,India[2]

ABSTRACT: With the advent of increasingly versatile information growing every day, more and more methods are used to facilitate the growing information. This information (also known as big data), is the very essence of our technological advancement, but immense care must be taken to maintain the security and privacy that should go hand in hand while processing this information.

In this paper, we shall look into the very definition of big data, how it is stored in a cloud system, the issues regarding security and privacy of the big data stored and what measures have been taken so far to overcome these problems.

KEYWORDS: Big Data; Biometrics; Hadoop; Privacy; Security; Twitter Data.

I. Introduction

Big data is the term given to any data that is too large to be stored by conventional methods. With the sheer size (volume) going in Exabyte and the diversity (variety) of information that big data holds, there is a great need to store this information in an environment beyond the conventional data storage units. This is where the need for cloud computing arises. Cloud computing is a revolutionary method of storing big data on the servers and keeps the authenticity (veracity) of the big data. However, big data did not have the idea of security nor privacy [1] during its conception. With the exponential growth and usage of big data, the need for security has never been greater. In this way, more and more methods are used to find out safe, efficient and cost effective ways of handling this situation.

As the big data is based on the 5 V’s, the issues [2] related to the privacy and security of these factors are as follows: Fig 1 shows the 5 V’s of Big Data and how each unit contributes to the entire big data as a whole.

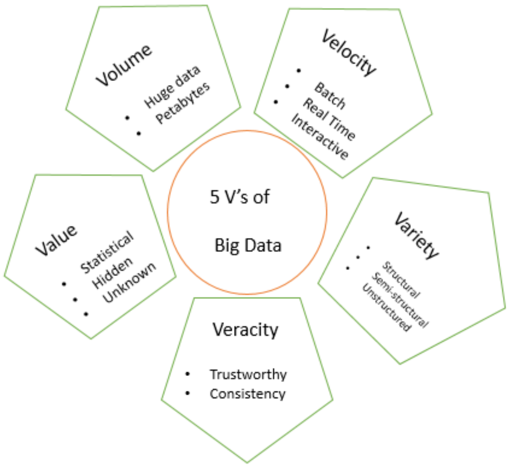
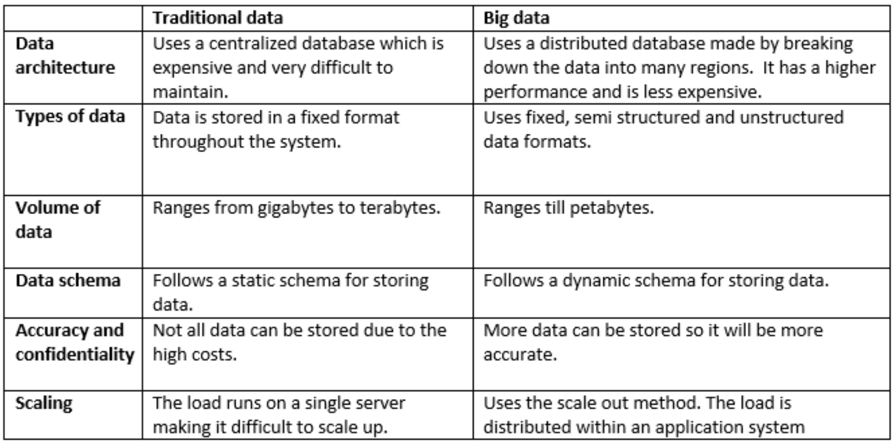


Fig 1 shows the 5 V’s of Big Data.

1. Volume: The sheer size of the big data makes it a humungous task to sort out and enable security and privacy features for each and every unit of the big data.
2. Velocity – Since the big data represents real-time values, there will be difficulty maintaining security measures on all the data that is transferred to and from the storage unit.
3. Veracity – If data is tampered with, it undermines the authenticity of the given system. This means that there will be a lot of doubt of the integrity of the gathered data.
4. Value – The data itself holds some degree of importance which can be breached and tampered with if it gets into the wrong users.
5. Variety – Structured or unstructured data shows great challenges to the diversity shown by big data in the matter of privacy.

When looking at how dissimilar big data from the earlier versions of traditional data, this table (Table I) shows the main differences.



	Traditional data	Big data
Data architecture	Uses a centralized database which is expensive and very difficult to maintain.	Uses a distributed database made by breaking down the data into many regions. It has a higher performance and is less expensive.
Types of data	Data is stored in a fixed format throughout the system.	Uses fixed, semi structured and unstructured data formats.
Volume of data	Ranges from gigabytes to terabytes.	Ranges till petabytes.
Data schema	Follows a static schema for storing data.	Follows a dynamic schema for storing data.
Accuracy and confidentiality	Not all data can be stored due to the high costs.	More data can be stored so it will be more accurate.
Scaling	The load runs on a single server making it difficult to scale up.	Uses the scale out method. The load is distributed within an application system

Table I shows the difference between traditional data and Big data

Thus privacy handling is critical to the development of big data. This means that there should be some effective methods that have to be placed. As of now, since the majority of the big data is stored in the cloud systems, there should be some form of measures that these systems should place to better the privacy and security of the system.

Cloud computing and security are closely linked with each other, this brings the concept of cloud computing security [3] to light. Cloud security enables the system to protect itself from the vulnerabilities faced from handling big data. In this way, modern computational features have enabled it to be better than before though there are still a variety of discrepancies.

II. Related work

Paper [2] has proposed various security and privacy challenges related to big data. Besides these challenges, they have also proposed techniques and methods to protect this data.Paper [4] talks about how biometrics can be used to secure big data. The applications and potential for biometrics in big data are discussed here. Paper [5] shows an application of big data in social media. Twitter data is a variation of big data as the amount of information that this site obtains on a daily basis is a huge feat by itself. Paper [6] talks about how big data is vulnerable to the outside world and how cloud computing and Hadoop have taken strides to combat this problem. Paper [11] proposes a meta cloud storage architecture which can be used to encrypt the Big data in the cloud. Paper [16] and Paper [17] discusses the various issues related to big data.

A. *Analyzing Big Data:*
Big data by itself is raw, untapped information. So care must be taken to ensure that this does not get into the hands of those who would misuse this data.

i) Security and Privacy Challenges
Some of the various problems faced while looking at the scope of big data is as follows.

1. Random distribution
Since big data is based on the concept of parallel computing [3], the issue lies in the fact that the data is stored in a random method and this leads to security issues. There is a need for regulating how the storage is distributed.
2. Privacy
Since big data assigns all value at the same priority, sensitive information is vulnerable to impeding attacks to these clusters.
3. Computations
The computations should be hidden as a way of protection from systems attempting to spy and extract these results.
4. Integrity
Since data is obtained at a large scale, there has to be special care taken to check for the validity and trueness [2] of the big data to prevent relying on compromised records.
5. Communication
Since there is no security between the inter node clusters that work when communicating with each other during the communication of big data, some secure network protocols [7] should be kept in mind so as to protect the interactions that take place among the various parties.
6. Access control
There is a need for a strong access control system that should be present in order to prevent any unwanted parties from getting access to the storage events of big data. Also, those clusters [8] which are under modification of any sort should be monitored using an authentication mechanism which prevents the system from getting attacked by malicious nodes.

ii) Possible Techniques to Protect Privacy in Big Data
Since big data can be protected, the following are the proposed methods that can be used to protect how big data is processed within the system.

1. Rules and Legality
Since governments and organizations collect and store all the digital records, it is bound to store sensitive information. In this way, since big data has a worldwide distributed storage, regions that uses this big data should be secure and comply with the rules that have been imposed on the system. In this way, a set of laws and legal rules [7] should be developed in order to make big data safe and beneficial to the system.
2. Encryption
Encryption can be used for the following factors:

- Storage
  Since big data is stored in clusters without any prioritization between sensitive and non-sensitive data, if a malicious party gets access to this unit, sensitive data can be easily tampered with. In this way, either all data or just sensitive data should be encrypted to protect the privacy of the sensitive information.
- Computations
  The computations define how the big data is being processed. Although it is difficult to predict where and how the big data is processed in a given time, some control schemes should be implemented so as to check and if necessary, deny nodes from accessing the results. To achieve this, there must be a method of executing a set of instructions without noticing the natures. This can be done through blind processing techniques based on holomorphic encryption of big data.

- Communications

Protocols such as SSL, TLS or even IPsec protocols [7] which make the communications unreadable without the knowledge of the keys are a useful way of securing different exchanges between the parties of big data environments.

3. Authentication

The big data architecture has to find new ways of controlling both joining clusters and accessing critical storage as a means of authenticating the resources.

4. Meta Data and Tagged Data

One way of segregating the collected data based on importance is through the involvement of metadata and tagging techniques. This will separate the sensitive data from the on sensitive and this will be able to maintain the necessary level of security when handling private information.

5. Unstructured distribution

To make the big data difficult to use by malicious parties that manage to hack the system, unstructured data makes it harder for them to steal any useful information. Unstructured distribution is also a good way of separating data [9] from related information and prevents hackers from extracting useful information if they ever have access to any nodes.

6. Tracing Activity

It is necessary to log every event that has been done over the big data and as well as the user who performed this. In this way, big data handlers will know where there was any malicious tampering of the big data.

B. *Biometrics in Big Data:*

The role of using physical features has been long used before. From identity tags to registration, it is very common to associate identification with a part of the body. Nowadays, technological advancements have made it in such a way that it is easy to mark unique characteristics of each and every person. This users a new path for biometrics [4].

i) Role of Biometrics in Authentication

- Fingerprint technology

  Since each fingerprint is unique, the scan takes into account the minute details of each fingerprint by analysing the intricate points of the given user. The finger points are noticed and the finer details are extracted by the machine's algorithm to give the se pattern for the finger.

- Palm recognition

  This is a step above the fingerprint scan as this analyses, the ridge features of the entire palm and then matches this data with similar patterns it has found within the database before giving a match. Since palms are unique and durable, they were known to be a reliable form of detection. Though it is stalling as there is some self-controls in the calculation and live scan technology.

- Facial recognition

  This captures the nodal points (up to 80) on a human face. These nodal points are the end points that are used to analyse the differences in a human face. Each differs in different humans based on the size or structure of the nose, cheekbones or even the depth of the eye socket.

- Iris scan

  Since the iris is unique to every person and as well as being unique within the same person (the left eye and right eye of a selected person will have a dissimilar iris). In this way, an iris pattern recognition will be able to be used for identification [14] and verification purposes.

The biometric approach is further analysed [15] with a series of mapping techniques as shown through the following figure. (Fig 2)

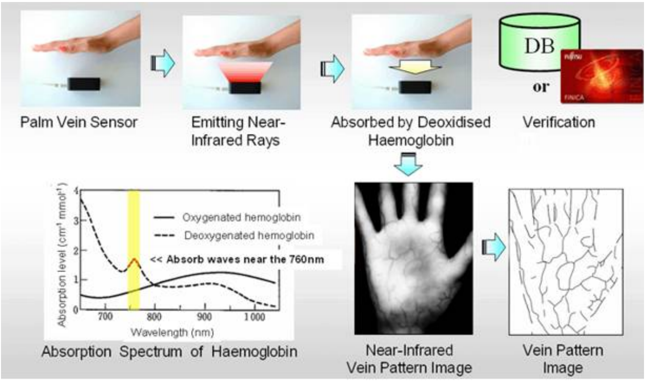

Palm Vein Sensor
Emitting Near-Infrared Rays
Absorbed by Deoxidised Haemoglobin
DB
or
Verification
Absorption level (cm⁻¹ mmol⁻¹)
4
3
2
1
0
700
800
900
1 000
Wavelength (nm)
Oxygenated hemoglobin
Deoxygenated hemoglobin
<< Absorb waves near the 760nm
Absorption Spectrum of Haemoglobin
Near-Infrared Vein Pattern Image
Vein Pattern Image

Fig 2: Palm Recognition

ii) Advantages of Biometrics in Big Data

- Biometrics cannot be replicated nor stolen. This will ensure better protection as compared to the standard verification methods.
- There is a diversity of biometric techniques such as face and iris scan that gives a more secure authentication services.
- This type of technology is user friendly, cost effective, less time consuming and at the same time, more reliable at authenticating a variety of users.
- This is easy to implement and at the same time, reduces the costs of issuing new identification methods for users who have lost their identification passwords or tokens.
- Biometrics is easily scalable.
- It is a very versatile security system.

C. Twitter Data Analysis through Supervised Classifiers:

In recent years, Social media has been one of the main sources for big data. A variety of data gets accumulated in the servers. If this data is analysed adequately then it provides us enormous information .Paper [5] proposes one of the methods that can be used to analyse twitter data.

i) Pre-processing

Pre-processing relies on obtaining information through the social media by finding related content and grouping it together to form meaningful information.

1. Collection of data

Obtained data from Twitter API and called it as Twitter4j using net beans. The data was searched using the #Hashtag and the movie name, actor, music record studios and production companies. This keeps the content limited to 140 words.

2. Normalization

Since user language has unnecessary values to the data system, all the tweets should be properly processed. The things that made the twitter data unclean are:

- URL’s: These distract the results that the classifier searches for as they are unnecessary.
- Username: There might be a way of adding a false user to the data and this should be cleared.
- Repeated characters: A character when repeated more than two times gives rise to a new word. In this way, these words should be removed from the system.

DOI: 10.15680/IJIRCCE.2017. 05032300

• Repeated words: If the word has appeared more than two times simultaneously, this should be converted to two words only.

3. Removal of stop words

The words such as ‘is’, ‘the’, etc. do not show the emotion of the user and are thus discarded.Once features are extracted the set of tweets of different categories were used to train the database and Naive Bayes classifier was used to classify the results into positive, negative and neutral tweets.

D. Big Data Hadoop Concurrent Processing:

Since big data is being used throughout various system, it is only natural that an in-depth analysis would lead to loopholes within the security and privacy. Paper [5] explains the role of this in everyday life.

i) Big Data Analysis

• Analysing a big data gives business models useful ideas that would help in growth, reforming unproductive work and show the characteristic strengths and weaknesses. Nowadays big data is being analysed and stored using Cloud services. This is very important and security should be a priority.

• Due to the very size of big data encryption is costly. Encryption is an important factor to prevent vulnerabilities that could result in loss of data. Hence encryption and decryption will influence the overall speed of observing and implementing the data. Multiple data types within a data set may be accessed and handled differently. In this way, it is difficult to track the access and data flow due to the high velocity of data with no standard part.

• Without the help of a proper data flow control, there will be risks involved when this content is shared among the cloud systems. Corruption in one data set will have the tendency to spread its corrupted data among other systems and thus putting the entire Cloud architecture at risk.

Paper [11] proposes a MetaCloud Data Storage Architecture to protect the big data against intruder in the cloud. As big data is by itself huge encrypting the entire data becomes impossible. So according to this method data is stored in multiple datacentres. To begin with the data is initially classified into three levels namely Sensitive, Critical and Normal which are then stored in different data centres. The proposed approach can redirect the user request to appropriate data centre in the cloud provided by different vendors. When the data is stored in the datacentre it will form a unique path. Thus instead of encrypting the entire big data the storage path is encrypted which will give a cryptographic value. The main objective of this architecture is to protect mapping of various data elements to each storage provider.

ii) Hadoop

Hadoop in an open-source framework used for Big Data analysis. Since Hadoop originally did not have security in mind during its inception [12], “Project Rhino” was unveiled by Intel by giving support for various encryption and authentication systems. These steps are not enough to stop cyber-crime [13] due to the large and important volume that is transmitted through. Hadoop architecture lets a file to be divided into chunks which are then transferred through various nodes which are concurrently processed. However these nodes can be extracted through chunk stealing and chunk injection. This threatens the security of the big data.One of the security advancements comes from a cloud communication company known as Twillio. This uses a highly reliable Hadoop framework using the Amazon S3 facility. Twillio implements a multi-tenant communications system to make sure that the isolation of rights of resources is followed. This is used to classify critical data and implement important security checks.

E. Issues related to Big Data and Hadoop

Big Data consists of wide variety of data. As the data is spread across the network and stored in distributed databases, it gives rise to various issues. Some of the issues are as follows-

Data Management Issues: This issue mainly focuses on maintaining the quality of data [16], its ownership, accessibility of data and documentation.

Data Storage and Processing Issues: The data is primarily stored using virtualization. Non SQL allows data to be stored in a non - tabular manner unlike the relational databases. Other tools include Apache Drill, Grid Gain and Hadoop. The data is processed in batch or stream processing. Map reduce is one of the processing techniques for big data [16].

1. Analysis Issues: The cookies incorporated in the websites and other apps installed on the devices collect personal information of the customer. However customers are unaware about the kind of analysis that is made on their data. Obviously the data is collected to predict the human behavior. For example when a person surfs the net, all related searches are displayed, also his activity is stored with the help of cookies .This information is then used to predict his interest. For example You Tube Likes can help predict the personality of a person, his interest etc. Thus there is a high tendency of over analysis of such data[17]. For Example if a person frequently searches for some information on a disease, then firms may use this information and market various health policies predicting that he is exposed to certain disease.
2. The HDFS Issue: The base layer of the Hadoop Architecture is the Hadoop Distributed Filesystem and it is the layer that is more vulnerable to security issues[16]. One of the method proposed to secure the data in Hadoop is to use Kerberos Mechanism where in a client is allowed to access a data node if and only if appropriate tokens are issued to him[16]

III. PRINCIPLE FINDINGS FROM THE SURVEY

The table below (Table II) shows how each big data factorscontributesto the overall development of the given system and how it still needs more improvements.

Table IIshows the principle findings from the survey along with their advantages and disadvantages

Paper	Big data factor	Parameter considered	Advantages	Disadvantages
[2]	Security	Rules and legal actions	Sets a standard for all big data architectures to follow.	Does not have a scope for those who find a loophole within the legal terms.
[4]	Biometrics	Facial Recognition	More secure at a personal level.	Changes done on the body will result in the system to consider the owner as an intruder.
[5]	Twitter data	Hashtag analysis	Can easily collect related work together to give valuable related information.	Some hashtag may have no relation to the data and might affect the data.
[6]	Hadoop	Infrastructure	Uses a Denial of Service to prevent outside attacks that could be risky.	The extra security measures compromises on computing speed and efficiency.
[11]	Security in cloud	MetaCloud Data Storage Architecture	The data is split and stored in different data centres and the storage path is encrypted.	The storage index for each data part has to be well maintained.
[17]	Security in HDFS	Kerberos Mechanism	Client access the data nodes only with the help of tokens. If the attacker steals the token then it can be renewed.	New users can only login if the server is always available.

IV. CONCLUSION AND FUTURE WORK

With so many applications of big data, it is very easy to see why this would be a target for a security breach. We are a long way from making big data a hundred percent secure. However, we have come a long way in making big data safe and secure to the best of our abilities. We can only hope that the future generations can transform the technology we have today to better this. With big data now taking a forefront in the growing Internet of Things, it is necessary to keep this data safe and secure.

REFERENCES

[1] Sangeeta, KapilSharma,Quality Issues with Big data Analytics, 'International Conference on Computing for Sustainable Global Development (INDIACom)', pp.3589-3591,2016.
[2] Youssef Gahi, MouhcineGuennoun, Hussein T. Mouftah, 'Big Data Analytics: Security and Privacy Challenges',IEEE Symposium on Computers and Communication (ISCC),978-1-5090-0679-3, 2016
[3] Peng-yu Wang and Ming-quan Hong, 'A Secure Management Scheme Designed in Cloud Security', IEEE 2nd International Conference on Big Data Security on Cloud, IEEE International Conference on High Performance and Smart Computing, IEEE International Conference on Intelligent Data and Security, pp.158-162, 2016
[4] VikashYadav, Monika Verma and Vandana Dixit Kaushik,'A Biometric approach to Secure Big Data', 1st International Conference on Innovation and Challenges in Cyber Security (ICICCS 2016),pp.75-79,2016
[5] RohitJoshi and RajkumarTekchandani, 'Comparative Analysis of Twitter Data Using Supervised Classifiers', International Conference on Inventive Computational Technologies(ICICT),Vol3,2016
[6] Ather Sharif, Sarah Cooney, Shengqi Gong, Drew Vitek, 'Current Security Threats and Prevention Measures Relating to Cloud Services, Hadoop Concurrent Processing, and Big Data', IEEE International Conference on Big Data (Big Data),pp.1865-1870,2015
[7] Umar Ahsan and Abdul Bais, ' A Review on Big Data Analysis and Internet of Things', IEEE 13th International Conference on Mobile Ad Hoc and Sensor Systems,pp.325-330,2016
[8] AdityaDev Mishra and Youddha Beer Singh, 'Big Data Analytics for Security and Privacy Challenges', International Conference on Computing, Communication and Automation (ICCCA2016) ,pp.50-53,2016
[9] Wang Jia, 'Study on Network Information Security Based on Big Data',9th International Conference on Measuring Technology and Mechatronics Automation,pp.408-409,2017
[10] AtifNaseer, Basem Y. Alkazemi and EhsanUllahWaraich 'A Big Data Approach for Proactive Healthcare Monitoring of Chronic Patients', Eighth International Conference on ubiquitous and Future Networks(ICUFN), pp.943-945,2016
[11] GunasekaranManogaran, ChanduThota, M Vijay Kumar, 'MetaCloudDataStorage Architecture for Big Data Security in Cloud Computing' 4th International Conference on Recent Trends in Computer Science and Engineering,Procedia Computer Science 87 (2016) 128 – 133
[12] Mike Uschold and Michael Gruninger, Ontologies: principles, methods and applications, 'The Knowledge Engineering Review', Vol. 11:2,93-136,1996.
[13] O. O'Malley, K. Zhang, S. Radia, R. Marti, and C. Harrell, 'Hadoop security design', Yahoo, Inc., Tech. Rep, 2009.
[14] http://searchsecurity.techtarget.com/definition/biometrics
[15] http://www.palmsure.com/technology.html
[16]B. Saraladevi, N. Pazhaniraja, P. Victer Paul, M.S. SaleemBasha, P. Dhavachelvan, ' Big Data and Hadoop-A Study in Security Perspective' 2nd International Symposium on Big Data and Cloud Computing (ISBCC'15),Procedia Computer Science 50 (2015) 596 – 601
[17]KshetriN,'Big data's impact on privacy, security and consumer welfare',Telecommunications Policy38(2014) 1134–1145

BIOGRAPHY

Shawn Rahul D'Souza is currently studying his Bachelor of Engineering in Computer Science and Engineering at St. Joseph Engineering College, Vamanjoor,Mangaluru, India. With an interest in the Internet of Things, he has pursued a paper to one of the fastest growing fields in order to learn how this data is stored.

Nisha Jenifer Roche is currently working as Assistant Professor in St Joseph Engineering College,Vamanjoor, Mangaluru,India. She received her Master of Technology(Mtech) in Computer Science and Engineering from NitteMeenakshi Institute of Technology,Bengaluru, India.Her area of research interests includesBigData, Digital Image Processing, and Internet of Things.