VOL. 10, NO. 5, MARCH 2015 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences


www.arpnjournals.com

EFFECT OF BIG DATA CHARACTERISTICS ON SECURITY-LEVERAGING EXISTING SECURITY MECHANISMS FOR PROTECTION

K. V. S. N. Rama Rao[1, 2], M. Pranava[2] and A. Mounika[3]
[1]Department of CSE, MLR Institute of Technology, Hyderabad, India
[2]Jawaharlal Nehru Institute of Advanced Studies (JNIAS), Hyderabad, India
[3]Department of IT, MLR Institute of Technology, Hyderabad, India
E-Mail: kvsnramarao@yahoo.co.in

ABSTRACT

Big Data is the surge of data which was caused by growing technology and the increase in online computing. Several characteristics that define Big Data are Volume, Velocity and Variety. The inherent nature of these characteristics will certainly introduce several vulnerabilities and threats for the entire data. These security concerns in Big Data must be addressed. The traditional systems security concerns were addressed by several strong security mechanisms. All these mechanisms are proved to be efficient and well functioning. In this paper, we discuss about the security issues that arise due to Big Data characteristics like Volume and Variety. Further, we focus on leveraging existing security mechanisms to overcome the effects of Big Data characteristics

Keywords: big data, security, big data characteristics, security mechanisms.

1. INTRODUCTION

There is a tremendous increase in technology for the past few years. This growth was accompanied with enormous amount of digital data; the Terabytes of data has increased to Petabytes and Zettabytes of data. The data is increasing exponentially everyday without any limit. This surge of data is known as the Big Data. The world around us is converting from manual to online world, where everything is being done online. Every interaction and communication being done online has become a source of data. This change has resulted in the advent of Big Data.

Big Data encompasses a wide range of data like Business exchange, stock market, organization logs, hospital records, research data, personal information, mobile records, geographical data, CCTV recordings, etc. These fields produce large volumes of data at a very high speed. The data produced is inconsistent and varied in nature.

As the data is growing, the issues related to Big Data are also increasing. Several Big data issues to be addressed are analysis, data storage, data management, etc, emerged. The major issue to be dealt with Big Data is Security and Privacy of Big Data. This concern primarily arises due to the inherent nature of Big Data characteristics. There is a strong need to address the effects that are caused by Big Data characteristics. Since there are proved existing security mechanisms, this paper focuses on leveraging certain existing security techniques to overcome the effects of Big Data characteristics.

The remaining part of the paper is organized as follows: In section two, we present the big data Characteristics. In section three, we briefly outline the existing security mechanisms. In section four, we present the effects of Big data characteristics and in section five we outline what existing security mechanisms can be leveraged. In the next section, we briefly conclude.

2. BIG DATA CHARACTERISTICS

META Group analyst Doug Laney has defined three-dimensions to describe data (known as the 3 V's), which are used to define the Big Data characteristics. The characteristics that define Big Data are:

A. Volume

This characteristic describes the quantity of data. The data magnitude in the era of big data has crossed Terabytes and Petabytes is becoming metric to analyze data sets. The volume of Big Data is increasing exponentially day by day at an uncontrollable rate.

B. Variety

This dimension describes the wide range of data formats encompassed by Big Data. Big Data consists of diverse data forms like images, audio, videos, text documents, raw fact, mathematical data, email, etc. Big Data comprises of data from different fields like Business exchange, data from Health and Life science, Organization data, Research data, government records, Geographical data, data collected by Astronomers, etc.

C. Velocity

This dimension describes the data generation speed; the speed at which the data is being generated. Advanced technology has given rise to high speed internet which is in turn producing data at a very high speed. Every day, internet users across the globe are producing data with every click of their mouse on web browsers.

Many communities, groups or organizations make use of additional dimensions to describe the Big Data characteristics such as Variability, Veracity, Complexity, Volatility and Validity.

D. Variability

This dimension describes the inconsistency in the data inflow at a particular time. That is, the amount of data produced at a particular time is very irregular. Sometimes the data production may be high and other times very low. The unpredictable behaviour of the data is responsible for many concerns in the Big Data area.

E. Veracity

This dimension describes the quality of data stored in the Big Data environment and its understandability. Data produced globally has different qualities. Many Big Data users are vary of the data they use and do not trust the information to make decisions.

F. Complexity

This dimension describes the complicated management of Big Data. The vast amount of Big Data has to be stored and analysed. The sheer size of Big Data makes these processes complicated.

G. Volatility

This dimension describes how long a data is relevant and for how much period of time the data should be stored. Sometimes a data is valid for a period of time and becomes invalid after that.

H. Validity

This dimension describes the correctness of the data. The storage of incorrect or wrong data leads to wastage of space. Hence this characteristic is used to describe the accuracy of Big Data.

These characteristics offered several differences between traditional systems and Big Data in terms of Data storage, Data sources, volume, variety, scalability, security, normalization, data redundancy and analysis of data. Since the present study scope is security, we discuss the security solutions for traditional systems.

3. SECURITY SOLUTIONS FOR TRADITIONAL SYSTEMS

Traditional system users make use of various security mechanisms to ensure safety of data. Some of the different mechanisms used in traditional systems are:

- **Access control:** Only the required and agreeable data is presented to the user.
- **Auditing:** The data is checked and verified periodically. Auditing is the process of regular inspection of data.
- **Authentication:** The user identification is verified. The common authenticity practice is login, where a user is asked to provide user ID and password.
- **Encryption:** Encryption is the process of converting the user data into an obscure form of data. The resultant data is not understandable to non-authorized user. Here a key is used by the user to convert the plain text (original data) into cryptic format (a non-understandable form of data). Only a user with the knowledge of the key used can decrypt the data (converting cryptic data into plain text) and read it.
- **Logging:** Logging is the process of recording every data entry. It is the process of recording data as it is produced. Logging ensures the availability of data. Any loss of data can easily be identified through logs.
- **Steganography:** Steganography is the process of hiding data behind or within another data. Sometimes data is hidden within pictures or images. This form of security is very effective.
- **Mandatory access control (MAC):** MAC is a type of access control in which the operating system restricts the access to system resources.
- **Validating and filtering:** Validating is the process of checking the correctness or truthfulness of data. And filtering is the process of removing unwanted data.
- **Syslog:** Syslog is a norm used in logging messages; it gives permission to detach the message generation from the software's that stores, reports and analyzes the messages.
- Homomorphic encryption is a process of performing operations on encrypted data as if it is a plain data without disturbing its cipher nature.
- Data integrity is a term used to refer to the data consistency and its correctness.
- **Auto-tiering:** It is the process of automated storage of data onto tiered storage devices.
- **Trust establishment:** In this process, the user is authenticated and their trustworthiness is proved.
- **Security information and event management (SIEM):**
- In SIEM technology, the security alerts generated by variety data is analysed. Its main focus is collecting, analyzing and represent variety data.
- **Dynamic provable data possession (DPDP):** In DPDP the updates of stored data are pre-processed and then it is stored onto untrusted servers. A small amount of metadata is held back. Later on, the correctness of data is asked to be proved.

The above security mechanisms are proved to be functioning effectively for traditional systems. Since Big data is exhibiting severe alarms with regard to security and privacy, a serious attention need to be given. In the next section, we briefly summarize the effects of Big Data characteristics such as volume, variety etc on security.

4. EFFECTS OF BIG DATA CHARACTERISTICS ON SECURITY

- To compute large volume of data in a distributed environment, framework like Map Reduce is used. The mapper reads data and reducer produces output. In some scenarios, the mapper could be untrustworthy and introduce malicious data content. Detecting the malicious content in such large volume of Data is huge task.

- In a cloud environment, the data logs are stored in different tiers. The advent of big data has initiated auto-tiering to store the large volume of Big Data. This will introduce security challenges like un-trusted storage services and no control over data storage location.
- Intrusion Detection in Big Data environment is a challenge due to the generation of several alerts for large data consecutively followed by many false positives. Handling such huge number of false positives in such a large volume of data is a challenge.
- Preserving the privacy will always top the agenda. The large volume of data that is collected will contain private information. An un-trusted person can get access to the entire volume of data easily and may compromise privacy. The volume of Big Data hinders in the process of privacy preservation, dynamic data operations etc.
- To obtain complete information regarding an attack, we need to perform auditing. But in the case of Big Data, the span and depth level measures of volume are enormous and cause major errors.
- The security attacks for the traditional systems are studied and a pattern is formed using the information. The process of pattern discovery, trends and correlations is important to implement security mechanism. For Big Data pattern analysis, the volume makes it a strenuous task.
- Organizations recruit experts for real time monitoring the security proceedings. This practise is possible for traditional systems as the data is comparatively low. Security monitoring in the case of Big Data is compromised due to its huge volume of data.
- Traditional systems maintain data and transactional logs. Manual storing of log gives experts grip over the data. Volume plays a spoil sport for manual logging in Big Data.
- Encrypting large volume of data is a time taking process and sometimes may even cause performance issues.
- Validating and filtering is a challenge in Big Data due to its volume. A user has the ability to produce malicious content and this data can be collected and stored.
- In the case of Big Data, the data will be collected from wide variety of sources. Many of these databases were designed to handle analytics issues but security is compromised when faced with Big Data variety.
- Metadata makes securing Big Data easier. The growing volume and variety complicates metadata provenance.
- Traditional systems implement steganography to protect mainly graphical data. The wide variety of Big Data comes across some adversaries while implementing steganography.
- Non-relational databases enforce security measures on middleware which does not provide strong security.

To address such severe effects caused by the characteristics of Big Data on security, there is a strong need for standard, proved and effectively functioning security mechanisms. In the section 3, we have summarized such security mechanisms. In the next section, we outline how these security mechanisms can be leveraged for Big Data security.

5. LEVERAGING EXISTING SECURITY MECHANISMS FOR PROTECTION

- To resolve the issue of untrusted mapper, two techniques can be implemented: trust establishment and Mandatory Access Control (MAC). In the trust establishment, the mapper is authenticated before assigning any task. In MAC process, various predefined security policies will be executed. An existing technique, Accountability Test (A-test) can be used to recognize the untrusted mapper. A-test identifies the malicious nodes by examining the working machines when a job is being executed, with the help of a group of trusted machines called as Auditor Group (AG).
- To address the challenges faced by auto-tiering in Big Data, strong solutions are not available. But feasible solutions can be implemented like a formal framework of Dynamic Provable Data Possession (DPDP), homomorphic encryption, privacy preserving auditing and network based auto-tiering.
- Big Data machine learning has to be implemented to discover anomalies and to handle with false positives. Recent study has revealed that Machine learning in Big Data is feasible and has been implemented in the analysis of Big Data in bioinformatics and applications like Siri in Apple iPhones.
- To reduce the risk of data exposure to un-trusted person, level of exposure should be reduced by implementation of access control at infrastructure level. In addition, authentication, reduced application complexity, encryption of data, protection of data sharing and compulsory access control should be strengthened.
- To provide successful auditing, techniques such as auto logging, forensic tools, Secure Information and Event Management (SIEM), and syslog on routers can be implemented.
- For a robust pattern discovery, efficient Big data machine learning can be implemented to discover the patterns.
- For real time monitoring issue in Big Data, research is being done and new tools and software's are being developed.
- To overcome the manual logging challenge, Auto-logging can be implemented which could remove the possible human errors.
- The use of cryptography to big data can be confined to attribute encryption. This will ensure protection to the data. For a successful Encryption, the cryptographic protocol should implement reduction

argument or simulation argument. An efficient and scalable Attribute Based Encruption (AEB) can be used to ensure protection of the most sensitive data.

- The Cloud Security Alliance has introduced three step processes to tackle the validation and filtering issue.

a) Improve the security of data collection platforms and applications
b) Determine possible attacks and reduce them.
c) Develop algorithms to detect and filter malicious data

- To address the issue in NoSQL, the Cloud Security Alliance suggested augment of middleware security policies and the security standards of NoSQL up to Relational Database. Combining data integrity with data encryption is an optimal way to provide security to NoSQL. Another way is to use NoSQL through frameworks. This will surround the database with a virtual layer and increase the security. Or a middleware can be placed around NoSQL and combining this with data encryption will increase security.
- The metadata provenance has to be secured in Big Data applications. Combining fast and lightweight authentication technique with provenance tools is key factor in securing data provenance and prevents outsider attacks. And for insider attack prevention, a dynamic and scalable access control is important.
- To scale the SIEM technology to respond to the Big Data issues, the organizations have to analyse their back-end volume and front-end capabilities. And implement SIEM which can accommodate the surge and dynamic analysis.
- New technologies are emerging for unstructured data analysis. The new approach makes use of scalable
- "share nothing" architecture, parallel databases, non-relational databases, and distributed processing frameworks.
- To implement steganography in Big Data, the data has to be broken down into manageable or known size using mathematical models and the private data has to be hidden behind these sizable data.
- To provide the necessary protection for non-relational databases, the augment of security policies for middleware should be done.

CONCLUSIONS

Big Data is the future of computing and its presence is increasing. But along with this growth, the inherent characteristics of Big Data such as volume, variety, etc are introducing several challenges. Since enterprises analyze and use this data for making key decisions, these challenges need to be addressed. One of the key challenges is the effect of Big Data characteristics on Security and Privacy. The Volume and Variety characteristics introduce several vulnerabilities and subsequently damage the data. We have described several security related issues which may arise due to the Big Data characteristics in this paper. To address these issues, utilizing existing standard and proved security mechanisms will be wise and ideal. Hence we have outlined several existing security mechanisms and their application to fight against the effect caused by Big Data characteristics i.e. we have discussed few possible measures for Big data security using existing security mechanisms and how they can be used, augmented or combined to provide feasible protection. Our future work aims in developing new techniques in combination with existing ones to enhance the security further.

REFERENCES

Trend Micro. 2012. Addressing Big Data security challenges: the right tools for smart protection, pp. 1-7.

Cloud Security alliance. 2012. Top ten Big data Security and Privacy Challenges, November, pp. 1-11

2013. Infosys Labs Briefings,"Big data Challenges and Opportunities", Vol. 11, No-1.

Sam Curry *et al*. 2013. Big Data fuels Intelligence driven security", January, pp. 1-12.

Venkata Narasimha Inkollu *et al*. 2014. Security issues associated with big data in cloud computing, IJNSA, Vol. 6, No.3, May, pp. 45-56.

2014. NIST. Big data Interoperability framework-Security and Privacy requirements, Vol. 4, pp. 1-39.

Schmitt, C., Shoffner, M., Owen P., Wang, X., Lamm, B., Mostafa, J., Barker, M., Krishnamurthy, A., Wilhelmsen, K., Ahalt, S., and Fecho, K. 2013. Security and Privacy in the Era of Big Data: The SMW, a Technological Solution to the Challenge of Data Leakage. RENCI, University of North Carolina at Chapel Hill. Text. http://dx.doi.org/10.7921/G0WD3XHT.

Zettaset. 2013. The Big data Security Gap: Protecting the hadoop cluster, pp. 1-8.

Tankard, Colin. 2012. Big data security. Elsevier Network security Journal, Vol., no. 7, pp: 5-8.

Omer Tene and Jules Polonetsky. 2012. Privacy in the age of Big Data, Symposium Issue, 64 STAN. L. REV. ONLINE. 63, pp. 63-69.

Steve Lohr. 2012. The Age of Big Data, February, pp. 1-4.

Zhifeng Xiao and Yang Xiao. 2011. Accountable MapReduce in Cloud Computing, The First Workshop on Security in Computers, Networking and Communications, pp. 1-6.