Unstructured Big Data Processing: Security Issues and Countermeasures

Shivasakthi Nadar, Narendra Gawai

Abstract - Every organization has big data. This big data contains structured, semi structured and unstructured data. Social networking users are increasing so the data of the social networking sites are also increasing rapidly. Mostly these data consists of images, videos, audios, conversations and e-mails. They are unstructured big data. So there is a need to process this data intelligently. This paper describes the security issues, challenges, vulnerabilities, attacks and the counter measures of Unstructured big data.

Index Terms - attacks, big data, countermeasures, processing, security issues, vulnerabilities, unstructured data.

1 INTRODUCTION

The Data volumes are increasing rapidly so processing such huge amount of data has become very difficult. Big data is defined as the five Vs that is volume, velocity, variety, value and veracity [1].

Data volume refers to the large amount of data that are generated every second that is around zeta bytes. So storage of big data is one of the challenge[1].

Data is streaming in at infinite speed and must be dealt with in a frequent period. To react quickly enough to deal with data velocity is a big challenge for most organizations [1]

Various formats of data are Structured, Unstructured text documents, email, video, audio, bank and financial transactions. Many organizations are still struggling to handle varieties of data [1].

Value is an important feature of the data defined by the added-value that the collected data can bring. So processing big data informative value is important [1].

Big Data veracity ensures that the data used are trusted, authentic and protected from unauthorized access and modification. The data must be secured from its collection, processing and storing on protected and trusted storage facilities [1].

1.1 Unstructured Big Data Processing

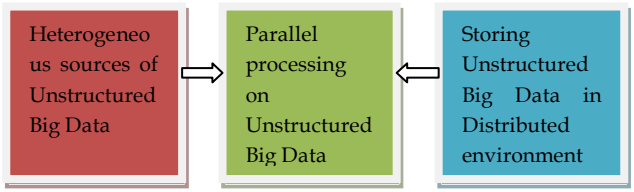

Fig. 1: Steps in Unstructured Big Data Processing

1.1.1) Capturing the Unstructured big data from heterogeneous sources

Unstructured Big data has to be captured from multiple heterogeneous sources. Particularly, due to popular online /mobile social networks, big data collection is of high volume, high velocity, and high variety. It was reported by IBM in Jan 2012 that 2.5 quintillion bytes of data are created every day, and 90 percent of the data (including structured, unstructured, streaming, and real-time data) in the world today were produced in the past two years. This presents new challenges, including how to efficiently store and organize high-volume data, how to quickly process streaming and real-time data, and how to accurately analyze unstructured data in order to maximize the value of big data [2].

1.1.2) Storing Unstructured Big Data in Distributed Environment

With advances in data storage technologies such as private and public clouds, now it is possible to store high-volume data. However, it is not efficient to move high volumes of collected data on the centralized storage. So distributed big data storage is suggested; that is, big data will be stored and organized in their original location [2].

1.1.3) Parallel Processing on Unstructured Big Data

Efficient processing of unstructured big data is essential for most of the organizations. Since big data are stored in a

distributed way, they should be processed in parallel either inter processing or intra processing so that new knowledge can be discovered. Compared to intra big data processing, inter big data processing is more challenging, as big data sharing should first be executed before processing, and during the time of data sharing, many new security and privacy issues arises[2].

2 Related Works

Advances in Big Data processing and value of privacy is described [3]. The challenges of Big Data with respect to Data storage, analysis [4]. A research methodologies for Big data analysis and design has been discussed [5].

3 Unstructured Big Data Security issues

Following are the security issues of Unstructured Big Data [6].
1) The big data is stored using Distributed architecture. So the data is partitioned horizontally, vertically, replicated and distributed among multiple nodes. So the data needs to be processed securely [6].

2) The unstructured data on social sites are changing continuously. So it is necessary to capture the changing data for processing.

3) There is a need to write the queries to handle the varying data.

4) The queries are the complex queries and needs to be processed in parallel.

5) Due to the large size of the data, instead of moving the data between the multiple nodes, it is feasible to move the code. So data security is essential.

6) Basically needs to store the large amount of unstructured data. So it is necessary to manage the large volume of data.

7) Due to its distributed architecture of big data storage, it is difficult to find out the exact location of the data among the available data nodes.
8) Big data captures the data from various logs, social media, etc. So we need to identify who has the right to access the data and at what time and from which location .

4 Challenges of Unstructured Big Data

4.1) Privacy and Security

It is the most important issue with Big data which is sensitive and includes conceptual, technical as well as legal significance. The personal information of a person after processing with external data sets generates new facts about that person. This information needs to be protected so it adds value to the business of the organization [7].

4.2) Data Access and Sharing of Information

If data is to be used to make accurate decisions on time so it becomes necessary that it should be timely available with consistency. Sharing the data of clients with other organizations needs to be secured [7].

4.3) Storage and Processing Issues

The storage that is available is not sufficient for storing bulk data which is being produced by the financial organizations and social medias. Sometimes this data needs to be uploaded on the cloud. Zeta bytes of data will take large amount of time to get uploaded in cloud and moreover this data is changing so rapidly which makes difficult to upload this changed data in real time.
The data transfer from storage to processing needs to maintain integrity. So it is necessary to build up the indexes in the beginning, for large amount of data processing [7].

5 Vulnerabilities in Unstructured Big Data

The following are the vulnerabilities found in Unstructured big data [6].

1) Redundant requests for computations in incorrect ways are performed on big data leading to capturing of the personal information, or credit card details and wrong values are stored in the database which leads to the loss to the financial organization and results into denial of service.

2) Due to the large set of data items, it is difficult to validate the input data and end point data which leads to wrong output.

3) To increase the performance, the access to the data has been given at the different levels which leads to the access of the data to the unauthorized person.

4) Due to the distributed environment of the big data from various sources such as real time data, social sites data, transactional data, it becomes difficult to secure this data.

5) Due to sharing of knowledge discovery from big data processing between the organizations, it becomes difficult to maintain the privacy of sensitive data.

6 ATTACKS ON UNSTRUCTURED BIG DATA

Attacks on an Unstructured Big Data is a major threat with respect to privacy of big data. There are three sub-categories for this type of attack, namely Equivalent database attacks, Human identification attacks, and specific identification attacks [8].

6.1) Equivalent database attack

While integrating the data from the similar databases, it leads to the leakage of database.

6.2) Human Identification Attack

When the data in one database is connected with the other database then human being (personal details) data gets captured .

6.3) Specific Identification Attacks

When we are searching for specific data in the database, at that time other associated private data gets exposed which leads to the privacy violation.

7 COUNTERMEASURES

Unstructured Big data security countermeasures recommendations with standard methods which can be effective for big data environments security [9]:

7.1) Authenticate nodes using Kerberos

It is effective to use Kerberos for validating inter-service communication and help keep unwanted nodes and applications out of the cluster. And it can help protect web console access, makes administrative functions tougher to compromise. Kerberos is one of the most effective security controls and it is useful to built into the Hadoop infrastructure [9].

7.2) Layer encryption for files

Encrypting a file protects against multiple attacker techniques from attacking application security controls. Encryption protects data if malicious users or administrators gain access to data nodes and directly inspect files, and renders stolen files or copied disk images unreadable. And file layer encryption provides consistent protection across different platforms regardless of OS/platform/storage type. It is transparent to both Hadoop and other applications, and scales out as the cluster grows. This is a cheaper way to address several data security threats[9].

7.3) Key management usage

Using key management service to distribute keys and certificates and manage different keys for each group, application, and user. This requires additional setup and possibly commercial key management products to scale with big data environment, which is bit complicated. Most of the encryption controls recommended depend on key/certificate security [9].

7.4) Create logs of activities

To detect attacks, diagnose failures, or to find out unusual behavior, you need a record of activity. Many web organizations start with big data specifically to manage log files. It gives a place to look when something fails [9].

7.5) Use of SSL/TLS

Implement Secure Socket Layer (SSL) or Transport Layer Security (TLS) between nodes, and between nodes and applications. Cloud era offers TLS, and some cloud providers offer secure communication, otherwise user need to integrate these services into their application stack [9].

8 CONCLUSION AND FUTURE WORK

This paper describes the processing, security issues, challenges, vulnerabilities, attacks on Unstructured Big data. We have suggested the countermeasures such as authenticate nodes using Kerberos, Layer encryption for files, Key management usage, creating log of activities and secure communication using SSL/TLS. Further the research work can be extended by implementing the unstructured big database by considering our suggested recommendations, to implement big data processing using Hadoop and MapReduce and also there is a scope for visualization analytics of unstructured big data process

REFERENCES

[1] Yuri Demchenko, Paola Grosso, Cees de Laat, Peter Membrey, "Addressing Big Data Issues in Scientific Data Infrastructure",IEEE 2014.

[2] Rongxing Lu, Hui Zhu, Ximeng Liu, Joseph K. Liu, and Jun Shao "Toward Efficient and Privacy-Preserving Computing in Big Data Era" IEEE Network , July/August 2014.

[3]Alvaro A. Cárdenas , Pratyusa K. Manadhata , Sreeranga P. Rajan "Data Analytics for Security", Co published by the IEEE Computer and Reliability Societies, Nov/Dec 2013.

[4] Jinsong Zhang, Yan Chen, Taoying Li College of Transportation Management Dalian Maritime University Dalian, China ,"Opportunities of Innovation under Challenges of Big Data", 2013 10th International Conference on Fuzzy Systems and Knowledge Discovery.

[5] Sachchidanand Singh, Nirmala Singh, "Big Data Analytics", IEEE, International Conference on Communication, Information & Computing Technology (ICCICT), Oct. 19-20, 2012.

[6] Marisa Paryasto, Andry Alamsyah, Budi Rahardjo, Kuspriyanto, "Big-Data Security Management Issues", 2nd International conference on Information and communication Technology 2014

[7] Avita Katal ,Mohammad Wazid , R H Goudar ,Department of CSE ,Graphic Era University "Big Data: Issues, Challenges, Tools and Good Practices", IEEE Conference 2013.

[8]Meiko Jenson "Challenges of Privacy Protection in Big Data Analytics" 2013 IEEE International Congress on Big Data.

[9] Securing Big Data: Security Recommendations for Hadoop and NoSQL Environments, Version 1.0 Released: October 12, 2012 , licensed by vormetric.

AUTHORS PROFILE

Narendra Gawai received his B.E. in Computer Engineering from Government College of Engineering, Amravati, Amravati University, M.E. from VJTI, Mumbai University, Maharashtra, India. He is currently working as Assistant Professor at UMIT, SNDT Women's University, Mumbai. He has 15 years of teaching experience. He has guided several undergraduate projects. His areas of interests are Systems Security, Digital Forensics, Big Data Analytics, Data Warehousing and Data Mining. Ph- 09503573013
Email id- ngawai@rediffmail.com

Shivasakthi Nadar received her B. Tech degree in Computer Science and Technology, from Usha Mittal Institute of Technology, SNDT Women's University, Mumbai. She is currently pursuing M. Tech 2nd Year from, Usha Mittal Institute of Technology, SNDT Women's University. Her areas of interests are Big Data Analytics, Image processing, Data Mining. Ph- 09869669349
Email id- shak_fire2008@yahoo.co.in