ISSN No. 0976-5697

Volume 6, No. 6, June - July 2015

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

Security Challenges in Big Data: Review

Vivekanand[1], Dr.B.M Vidyavathi[2]
M Tech 3rd sem[1], professor[2]
Department of CSE, BITM, Bellary, India

Abstract- Big data because of its various properties like volume, velocity, variety put forward many security challenges. Security of Big Data is a big concern. Big data phenomenon arises from the increasing number of data collected from various sources, including the internet. Since big data is a recent upcoming technology in the market which can bring more benefits to the business organizations, it becomes necessary that various security challenges and issues associated in bringing and making use of this technology are brought into light. Big data Hadoop tool is used for storage and processing the big data. During the initial development of Hadoop, security was not a prime focus area. This paper describes the apache hadoop, its present security mechanism, security challenges and survey of existing methods to handle security challenges.

Keywords:- Security, Big data, Apache Hadoop, Encryption

I. INTRODUCTION

Big data is currently a major topic across a number of fields, including management and marketing, scientific research, national security, government sector and open data. Both public and private sectors are making increasing use of big data analytics. The word big data refers to the large amounts of digital information companies and governments collect about us and our surrounding environments. Every day, we create 2.5 quintillion bytes of data so much that 90% of the data in the world today has been created in the last two years alone. And remaining 10% of the data has been created when those data storage systems is generated. In present world there are many data generalization factors or data resources those are sensors, CCTV cameras, social networks like Facebook, what's app, Gmail and many more. Online shopping's, airlines, hospitalists data from all these resources huge amount of data is being generated day by day, to handle these huge amount of data the big data is introduced.

The data which is beyond to the storage and processing power that data is usually known as a big data. Data is increasing at a huge speed making it difficult to handle such large amount of data (exabytes).The main difficulty is handling such large amount of data because the volume is increasing rapidly comparing to the computing resources. Security and privacy issues are increasing by velocity, volume, and variety of big data, such as large-scale cloud infrastructures, area of data sources and formats, streaming nature of data acquisition and high volume inter-cloud migration. The utilization of large scale cloud infrastructures, with a use of software platforms, spread across large networks of computers, also increases the attack zone of the entire system. The current growth rate in the amount of data collected is staggering. A challenge for IT researchers and practitioners is that this growth rate is fast exceeding our ability to both design appropriate systems to handle the data effectively and analyze it to extract relevant meaning for decision making and major challenge is with respect to security of the data.

The rest of the paper is organized as follows. Section 2 introduces Apache hadoop, its present security level and lists the security challenges. Then Section 3 explains about related work. And finally conclusion is added in Section 4.

II. APACHE HADOOP

For huge data storage and processing, initially google people were started working on web search engine in 1990 they have just given description or idea in white paper of Google File System (GFS) and MapReduce but they have not implemented. But yahoo people started working on the white paper which is published by google, finally in 2007 they have concluded with Hadoop distributed file system (HDFS) and in 2008 MapReduce. The combination of HDFS and MapReduce is known as a hadoop [2]. Hadoop is an open source frame work given by apache software foundation for storing and processing the huge data set with cluster of commodity hardware's. This framework is used by the plain programming models and which handles the distributed parallel programming models. Hadoop handles the size of the data sets are petabytes and exabytes across clusters of computers so that cluster of hadoop can easily scale out. Hadoop is made up of two components those are

a) Hadoop Distributed File System (HDFS)
b) MapReduce

a. HDFS:

Hadoop Distributed File System is a File System designed for storage of large files with streaming data access patterns, running on cluster of commodity hardware. HDFS block size is larger by default size is 64 MB compare to normal file systems. The reason for large size blocks is to reduce the number of disk usage and make use of total memory space. HDFS cluster has two types of nodes namenode (the master) and datanode (worker). The name node manages the file system namespace, handle the file system completely and the metadata for all the files and directories in the complete structure. Datanode stores and access the block, as per the instructions given by clients. The data retrieved is reported back to the namenode along with lists of blocks that they are storing in it. Without namenode it is not possible to access the file. So it is very important to make namenode is working properly.

b. MapReduce:

MapReduce is a method for processing huge data sets in parallel. The DataNode acts as compute node to have the computation nearby the data processing node. Each node has a TaskTracker which performs map and reduce on the submitted job. JobTracker schedules to job submit-ted by the user on certain compute node.

The authentication between the user and the JobTracker is done through Kerberos using RPC. A submitted job should run with the user identity and permission, as of now there exist no authentication mechanism for MapReduce than Service Level Agreement (SLA). MapReduce stores the information about the executing and remaining jobs in HDFS.

A. Present Hadoop Security Level:

Hadoop default means consider network as trusted and hadoop client uses local username. In default method, there is no encryption between hadoop and client host and in HDFS, all files are stored in clear text and controlled by a central server called NameNode. So, HDFS has no security appliance against storage servers that may peep at data content. Additionally, Hadoop and HDFS have no strong security model, in particular the communication between datanodes and between clients and datanodes is not encrypted. To solve these problems, some mechanisms have been added to Hadoop to maintain them. For instance, by strong authentication, hadoop is secured with Kerberos and thorough it, provides mutual authentication and protects against eavesdropping and replay attacks. Every user and service has a Kerberos "principal" and credentials are by Service: keytabs and User: password which RPC Encryption should be enabled.

Layers of defense for a hadoop cluster are [3]

a. Perimeter Level Security: Network Security firewalls, Apache Knox gateway
b. Authentication: Kerberos
c. Authorization: e.g. HDFS permissions, HDFS Access Control Lists, MR ACLs
d. OS Security and data protection: encryption of data in network and HDFS

According to survey conducted by Cloud Security Alliance members and security practitioner-oriented trade journals the list of high-priority security and privacy challenges, the following are the top ten security and privacy challenges [1] for big data and for hadoop tool.

a. Secure computations in distributed programming frameworks
b. Security best practices for non relational data stores
c. Secure data storage and transactions logs
d. End-point input validation/filtering
e. Real-time security/compliance monitoring
f. Scalable and composable privacy-preserving data mining and analytics
g. Cryptographically enforced access control and secure communication
h. Granular access control
i. Granular audits
j. Data provenance

III. RELATED WORK

Following are the existing methods to make Hadoop cluster more secure:

Security of hadoop cluster is implemented with apache sentry [5], is an open source project by Cloudera is an authorization module for Hadoop. This helps to achieve granular-based, role-based authorization to provide precise levels of access to the right users and applications. It takes care of role-based authorization, fine-grained authorization, and multi-tenant administration. Apache sentry key benefits include storage of sensitive data in hadoop and extend hadoop to more users. There is separate rules are maintained by separate admen's for each database or schema. Sentry is a major step taken in Hadoop security, making Big Data highly accessible by even more industries, organizations, and end-users and giving administrators the control of multi-tenant administration, and unified platform they need to make that happen easily.

Project Rhino [8] provides an integrated end to end data security solution to the Hadoop ecosystem. It provides a token based authentication and solution. It offers Hadoop crypto codec framework and crypto codec implementation for block level encryption on data stored in Hadoop. It supports key distribution and management so that MR can decrypt data block and execute the program as per instruction. It also enhances the security of HBase by offering cell level authentication and clear encryption for table stored in Hadoop. It supports audit logging framework for easy audit trails.

In Fully Homomorphic Encryption [7] method ensures the safety and reliability from the three levels of hardware data, users and operation level. Homomorphic Encryption technology enables the encrypted data to be operable to protect the security of the data and the accuracy of the application. Fully Homomorphic Encryption allows multiple users to work on encrypted data in an encrypted form with any operation. Currently because of the computational complexity, data increases seriously and for other reasons during use of fully homomorphic encryption, it has not been put into practical use.

A ciphertext policy attribute based encryption [9] method is based on access control scheme of cloud storage data security. This method is very easier to manage keys, but also more transparent to the user, this allows the users to take less time in the key generation and distribution also other works. To handle the security challenges of network and data sharing method in cloud storage service, this method proposes data security access scheme storage based on attribute group on cloud, so that data producers do not take part in the specific operation of the property. This method has high security and reliability, but the efficiency of the implementation has to be improved.

In triple encryption [10] method, it combines HDFS file encryption using Data Encryption Algorithm (DEA) and data key encryption with RSA. Later encrypts the users RSA private key using International Data Encryption Algorithm (IDEA). Principle of data hybrid encryption is that HDFS files are encrypted using a hybrid encryption method, a HDFS file is symmetrically encrypted by a single key k and the key k is then asymmetrically encrypted by owner's public key. Symmetrical encryption is safer and more expensive than asymmetrical encryption. This method planned to implement the parallel processing of the encryption and decryption using MapReduce, in order to improve the performance of data encryption and decryption.

IV. CONCLUSION

Big data security is the important issue for the big data developers. To handle big data, hadoop tool is introduced but this tool only focuses on storage and processing of big data. As big data is spreading faster like that security issues also growing faster. To handle these security issues, there are many other mechanisms and methods are implemented with hadoop to make hadoop tool as a more secure. In apache sentry method authorization is achieved only for few components of the hadoop, so other component of the hadoop required authorization that should be taken care. In project rhino token based authentication is used for each user, if token is stolen by third party then data loss will happen. In homomorphic encryption because of the computational complexity, data increases seriously and other reasons when using fully homomorphic encryption, it has not been practically implemented. Ciphertext policy attribute based encryption method has high security and reliability, but the efficiency of the implementation has to be improved. Triple encryption method fails to achieve parallel processing of encryption and decryption using MapReduce. From the state of art these methods it is found that still there are many security loopholes present in the processing and storage of the large data sets. Research to be carried out in this area is open.

V. REFERENCES

[1]. "Top Ten Big Data Security and Privacy Challenges" CLOUD SECURITY ALLIANCE https://cloudsecurityalliance.org/ November 2012.

[2]. Apache™ Hadoop®! Available: http://hadoop.apache.org/

[3]. Securing your Hadoop Infrastructure with Apache Knox. Available: http://hortonworks.com/hadoop-tutorial/securing-hadoop-infrastructure-apache-knox/, 2014.

[4]. S. V. a. B. Noland. With Sentry, Cloudera Fills Hadoop's Enterprise Security Gap. Available: http://blog.cloudera.com/blog/2013/07/with-sentry-cloudera-fills-hadoops-enterprise-security-gap, July 24, 2013.

[5]. X. Zhang, "Secure Your Hadoop Cluster with Apache Sentry," ed: Cloudera, April 07, 2014.

[6]. Securosis, "Securing Big Data: Security Recommendations for Hadoop and NoSQL Environments", October 12, 2012.

[7]. S. Jin, S. Yang, X. Zhu, and H. Yin, "Design of a Trusted File System Based on Hadoop," in Trustworthy Computing and Services, ed: Springer, pp. 673-680, 2013.

[8]. "Real-Time Big Data Analytics for the Enterprise", White Paper Intel® Distribution for Apache Hadoop Big Data, 2014.

[9]. H. Zhou and Q. Wen, "Data Security Accessing for HDFS Based on Attribute-Group in Cloud Computing," in International Conference on Logistics Engineering, Management and Computer Science (LEMCS 2014), 2014.

[10]. C. Yang, W. Lin, and M. Liu, "A Novel Triple Encryption Scheme for Hadoop-Based Cloud Data Security," in Emerging Intelligent Data and Web Technologies (EIDWT), 2013 Fourth International Conference on, pp. 437-442, 2013.

[11]. Yuri Demchenko, Zhiming Zhao, Paola Grosso, Adianto Wibisono, Cees de Laat, "Addressing Big Data Challenges for Scientific Data Infrastructure", IEEE , 4th International Conference on Cloud Computing Technology and Science, 2012.

[12]. Zettaset, "The Big Data Security Gap: Protecting the Hadoop Cluster", 2013. Available: http://www.zettaset.com.