Comprehensive Survey on Hadoop Security

Maria Martis, Namratha V. Pai, R. S. Pragathi, S. Rakshatha and Sunanda Dixit

Abstract The new emerging technologies have provided a way for a large amount of data generation. Secure storage of such a huge data is of prime importance. Hadoop is a tool used to store big data, where security of it is not assured. In this paper, we have considered a survey on various approaches which helps in providing secure storage of files in Hadoop. Hadoop framework is developed for the support of processing and storage of Bigdata in a distributed computing environment. Usage of Bigdata has become a key factor for the companies as they can increase their operating margin. Bigdata contains user-sensitive information and bring forth many privacy issues. Bigdata is a larger and a more complex datasets obtained from a variety of network resources. These datasets are beyond the ability of traditionally used data processing software to capture, manage, and process the data within the given time frame. These massive volumes of data are used by many of the organizations to tackle the problem that could not be done before. Since the data holds a lot of valuable information, these data need to be processed in short span of time by which companies can boost their scale and generate more revenue, traditional system resources are not sufficient for processing and storing, and this is where Hadoop comes into picture. The main objective of Hadoop is running of application of bigdata. Hadoop being a great tool for data processing, it was initially designed for internal use (i.e., within local cluster) without any security perimeter of organization, so they were easily hackable and exposed to threats.

Keywords Hadoop · Data security · Big data · Authentication · Authorization

1 Introduction

In the digital universe, due to the rapid growth, expansion of new services and number of online users on the Internet has lead to the significant increase in amount of data produced every millisecond. According to the Internet World Stats in the month

M. Martis (✉) · N. V. Pai · R. S. Pragathi · S. Rakshatha · S. Dixit
Department of ISE, Dayananda Sagar College of Engineering, Bangalore 560078, India
e-mail: mariamartis4@gmail.com

N. R. Shetty et al. (eds.), *Emerging Research in Computing, Information, Communication and Applications*, Advances in Intelligent Systems and Computing 906,
https://doi.org/10.1007/978-981-13-6001-5_17

of December 2017, there were 4.157 million data users, i.e., 54.4% of total world population [1]. An open-source framework named Hadoop is provided by Apache in order to handle enormous amount of data efficiently. Hadoop is a setup, and HDFS is a framework. Hadoop consists of name node and data node which will be master–slave configuration. The job trackers that run on name node accept the jobs and allocate it to the task tracker. Task tracker that runs on the data node is responsible for performing the task and reporting the completion of task to the job tracker. The input file is broken down into blocks by name node, and each of the blocks is allocated to the data node. Replicas are used to provide the fault tolerance. More than 50% of Big Data experiments are executed on Hadoop, so securing these data has become a necessity. Hadoop does not contain any security framework; a foreign user can easily attack the data and access the contents.

The Present Hadoop security level is as follows:

(1) There is no encryption technique between host and Hadoop.
(2) The files which are stored in text format are controlled by name node.
(3) There is no strong security for communication between data nodes and clients.

Some of the known general procedures to provide security are Apache Knox gateway, it provides single point authentication, and main protocols used in this are HTTP or HTTPs to Hadoop cluster that provide single access point, central authentication, authorization, and hides topologies. Next level is authentication which allows identifying who the user is. It is provided by Kerberos [2] which allows nodes communicating with each other to prove their identity among themselves in much secured manner. Authorization ensures that user can access only those data for which they are entitled. Knox gateway provides authorization by evaluating user, group, and IP address. Hadoop (Highly Archived Distributed Object Oriented Programming) Hadoop actually came into, the picture in 2005 is a package that offers documentation, source code, location awareness, work scheduling.

Hadoop addresses bigdata challenge [3] which is explained below.

1. Hadoop framework allows user to store bigdata in distributed environment where the data is stored in blocks across the data nodes and block size can also be specified.
2. Extra data nodes can be added to HDFS cluster; thereby, scaling problem is resolved.
3. The data written once can be read multiple times since there is no predumping schema validation; hence, a variety of data can be handled in Hadoop.

Hadoop being a great tool for data processing, it was initially designed for internal use (i.e., within local cluster) without any security perimeter of organization, so they were easily hackable and exposed to threats. Hence, encryption is used to protect the data stored in Hadoop, which is must in government and finance sectors worldwide to meet privacy and other security.

Hadoop is a efficient, reliable distributed platform that provides scalable storage and computing on large datasets. Hadoop offers framework to process a large volume of data by running a large number of jobs in parallel on a cluster of machines.

Organizations use Hadoop because of its ability to store and process huge amounts of data quickly. But when they started storing confidential sensitive data on Hadoop clusters, a need for strong security mechanisms to protect these data is observed.

The user data stored in the cloud is not a controllable domain, and in order to protect the important data of user, confidentiality is an issue of most concern. In traditional public encrypt mechanism, the encryption resource provider needs to obtain all relevant information of user, and it will damage the user's privacy and requires more bandwidth and large processing overhead. The CP-ABE method has been proposed to solve the above issue.

The 4 Modules of Hadoop:

- Distributed File System

The most important two are the distributed file system, which allows data to be stored in an easily accessible format, across a large number of linked storage devices, and the MapReduce, which provides the basic tools for poking around in the data. (A "file system" is the method used by a computer to store data, so it can be found and used. Normally, this is determined by the computer's operating system; however, a Hadoop system uses its own file system which sits "above" the file system of the host computer—meaning it can be accessed using any computer running any supported OS).

- MapReduce

MapReduce is named after the two basic operations this module carries out—reading data from the database, putting it into a format suitable for analysis (map), and performing mathematical operations, i.e., counting the number of males aged 30+ in a customer database (reduce).

- Hadoop Common

The other module is Hadoop Common, which provides the tools (in Java) needed for the user's computer systems (Windows, Unix, or whatever) to read data stored under the Hadoop file system.

- YARN

The final module is YARN, which manages resources of the systems storing the data and running the analysis. Various other procedures, libraries, or features have come to be considered part of the Hadoop "framework" over recent years.

2 Hadoop Security Methodologies

This section discusses various Hadoop security methodologies.

2.1 Data Leakage Detection Using Haddle Framework

Gao [4, 5] proposed Haddle, a forensic framework that helps us to find the illegally copied and offender who stole the data. Haddle can improve the inspection of financial records of Hadoop. Haddle uses data collector and data analyzer to collect Hadoop logs, Fsimage files are sent to server, and automatic analytic method is used to find stolen data and the offender to recreate the crime scene, respectively. Investigating a big Hadoop cluster to identify the attacked node is very difficult, but this can be achieved by using automatic detection algorithm and abnormal condition can be noticed. Hadoop progger provides a lot of evidences even if Hadoop logs or Fsimage is compromised. Better techniques for collecting files are to make data contingent and reduce the performance issues. More improvement must be done on attack detection algorithm.

2.2 CP-ABE Security Access Mechanism

Zhou [6] proposed CP-ABE; it is a type of encryption where both private key and encrypted text depend on attribute. The system includes users and the intersection occurs when user and the encrypted text which is to decrypted consists pack of attributes which is greater than or equal to threshold and its user-specific key helps in regaining the original text. The proposed method prevents obtaining user information and reduces the chances of violating user rights. The approach provides data security accessing in the Hadoop throughout an area. The efficiency of implementation is yet to be achieved.

2.3 Data-at-Rest Security Using SDFS

Petros Zerfos [7, 8] proposed a method in which data at rest is provided security by using Hadoop in the cloud service by developing a new Hadoop file system called secure distributed file system (SDFS). The performance bottleneck that arises from the key distribution lowers the storage requirement for the secure data by using secret sharing and information dispersal technique. End-to-end security and controlled access of the data stored in enterprise are provided by SDFS which is used to minimize computational overhead and cost.

2.4 Modified RC4 Technique Using MapReduce

Jayan [9] proposed a method which uses parallel sections and modified RC4 algorithm to encrypt the data. The input is split into blocks and applied the encryption algorithm. The output is combined to get text in non-readable form. The algorithm is made parallel by combining the keys with the number of threads. This enhances the security. The key scheduling part and the random number generation part are executed parallel. Because of the parallel generation of keys, modified RC4 algorithm shows better performance and is more cost-effective than the MapReduce algorithm. Although modified RC4 algorithm is parallel, it is not capable of performing according to the expected level.

2.5 Cloud Disk Security Based on the Hadoop

Jing [10] proposed cloud disk storage which upholds the confidentiality. The encryption is done by the symmetric encryption algorithm and RSA algorithm which are operated on Hadoop cluster, and cloud is used for storage of data to give security. Performance was evaluated by comparing the expenditure and the files loaded into storage. The security provided by the cloud is not efficient enough.

2.6 Data Security Based on Hash Chain Technique

Jung [11–13] proposed a system which introduces an additional hash chain with one-way hash function. The hash function h of the PK scheme is utilized, and the output of *h* is again hashed; it takes only one hash value; thus, two values are required for the operation. It provides improved performance compared to PK scheme. PK scheme is based on one-time token method and prevents vulnerability of the block access token, which acts as a proof of user access rights on the data blocks. If a block access token is exposed to an attacker during its usage, the token cannot be used in a replay attack. It also offers high fault tolerance and high availability.

2.7 Security in G-Hadoop

Jam [14, 15] proposed the technique G-Hadoop that runs on many collections in a grid working area. G-Hadoop uses Secure Shell protocol between the client and the Hadoop cluster, it uses Globus Security Infrastructure (GSI) for providing security. GSI is one of the standards which is used for the grid security. Setting up GSI architecture provides security to each Hadoop cluster. It provides single sign on process, many other authentication mechanisms and upholds the integrity of messages sent over the grid. With the help of SSL hand shaking mechanism, communication between job node and the data node is securely established. The above approach provides better scalability.

2.8 Fully Hamomorphic Encryption

Jin [15] proposed work uses two technologies, namely fully homomorphic encryption technology and authentication agent technology. Former allows many users to experiment on data in form of unreadable format with different operations and yields same result as the data had been unlocked. Homomorphic encryption technology encrypts data first and then stores in data storage and latter combines many access control mechanism. The main drawback is the large increase in data and high complex computation.

2.9 Hadoop Security in Public Cloud

Yu [12] proposed SEHadoop [16, 17] model; this is developed to make Hadoop recover quickly from any malicious attack. This can be achieved by isolating Hadoop components, limit the range of data that can be accessed and gives entry license for each process. Setup for security was made on name node resource manager Kerberos using SE block token. In SEHadoop, there is no over usage of authenticated key, and it provides access control with minimum expenditure. SEHadoop includes SEHadoop block and delegation token, where SEHadoop block token does not possess any heavy burden or depreciation, whereas SEHadoop delegation token shows very limited accomplishment impact on existing Hadoop.

2.10 OTP Authentication Technique

Somu [18] proposed one-time pad technique that prevents the transfer of password in between servers by using one-time pad algorithm. It involves two steps:

1. Registration process,
2. Authentication process.

In the registration process, user enters the username and password. The password is encrypted two times using the OTP algorithm and mod 26 operation and stored in the registration server. Once again the password is encrypted using the symmetric cipher technique and sent to the backend server along with the username. In authentication process, the user needs to enter the username, and the backend server sends the cipher to the user via the registration server. Registration server decrypts with the key returned by the user. The registration server compares username stored in the backend with the username entered by the user. If it matches, authentication of the user is successful. This technique helps to store the process data related to the credit cards and healthcare. User must register and obtain OTP as the security code to access the resource on Hadoop. The OTP is not much secure which makes the outsider to get it easily and get access to the individual system.

2.11 Hadoop-Based Cloud Data Security Using Triple Encryption Scheme

Chau Yang proposed a method which combines HDFS file encryption using DEA and data key encryption with RSA and encrypts the user RSA private key using IDEA. The hybrid encryption consists of choice against two forms of encryption. The proposed system uses DEA and RSA algorithm to encrypt the data and get the data key and then encrypts the data key. The private key is kept with the user in order to decrypt the data key. It increases the performance through parallel processing of encryption and decryption using Hadoop. The performance factor is calculated by comparing the speed of writing and file size which is given as input to the HDFS system. The future work aims to achieve the parallel processing using MapReduce framework.

Table 1 gives a detailed survey of various security methods for Hadoop.

3 Conclusions

In any storage environment, protection of data is of almost importance. The above-listed methodologies provide various levels of security on Hadoop. The above-listed methodologies do not provide security measure at a very large scale. The techniques specified in this paper yield the performance which is not feasible to the expected level. In future, there is a need of producing and implementing a variety of powerful security measure to tackle the problem of danger or threat for data loaded in Hadoop and let the enormous data to be in a state of feeling safe.

Table 1 Survey of different Hadoop security methodologies

Year	Title	Encryption/scheme	Methodology	Pros	Cons
2017	Modified RC4 technique using MapReduce	Modified RC4 algorithm	Parallel section and modified RC4 algorithm	Better performance and cost-effective than MapReduce	Not flexible as per the expected level
2015	Data leakage detection using Haddle	Automatic detection algorithm	Security is provided using data collector and data analyzer	Identifying abnormal condition in node can be determined	Enhancement of detection algorithm to analyze the real scene and data in efficient manner
2015	Data security based on Hash chain technique	Hash chain technique	Two hashing functions required	Improved performance compared to PK scheme	NA
2015	Hadoop security in public cloud	SE Hadoop model	Isolating Hadoop components and limiting range of data	Access control over minimum overhead	SEHadoop delegation token shows limited performance
2015	Data-at-rest security using SDFS	SDFS to secure data at rest by using Hadoop in cloud service	Secret sharing and informal dispersal	Minimal computational overhead and cost	NA
2014	CP-ABE security access mechanism	Encryption done based on attributes	Encryption done where both user-specific key and encrypted text depend on attribute	Prevents obtaining complete user information	Efficiency of implementation is yet to be achieved
2014	OTP authentication technique	One-time pad algorithm	Multiple password encryptions is done and sent to backend server	Efficiently used in credit card, health care, etc.	OTP is not much secure and cracked easily

(continued)

Table 1 (continued)

Year	Title	Encryption/scheme	Methodology	Pros	Cons
2014	Security in G-Hadoop	G-Hadoop security architecture	Setting up GSI architecture provided security to cluster	Provides authentication communication using asymmetric cryptography	NA
2014	Fully homomorphic encryption	Authentication agent technology	Encryption is done by fully homomorphic encryption and then stored in HDFS	Combines any access control mechanisms	Complex computation and high cost
2013	Hadoop-based cloud data security using triple encryption scheme	Triple encryption scheme	HDFS file encryption using IDEA, data encryption with RSA, RSA private key using IDEA	Increases performance through parallel processing	NA
2013	Cloud data security based on Hadoop	BANLOGIC	Symmetric encryption algorithm and RSA algorithm used to secure Hadoop cluster	Ability to provide flexible and low-cost services	Security was not sufficient enough

References

1. www.internetworldstats.com.
2. Park, S. H., & Jeong, I. R. (2013). A study on security improvement in Hadoop distributed file system based on Kerberos. *Journal of the Korea Institute of Information Security and Cryptology, 23*(5), 803–813.
3. Abouelmehdi, K., Beni-Hssane, A., Khaloufi, H., & Saadi, M. (2016). Big Data emerging issues: Hadoop security and privacy. In *2016 5th International Conference on Multimedia Computing and Systems (ICMCS).*
4. Gao, Y., Fu, X., Luo, B., Du, X., & Guizani, M. (2015). Haddle: A framework for investigating data leakage attacks in Hadoop. In *IEEE 2015*.
5. Chen, C. L. P., & Zhang, C. Y. (2014). Data intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences, 275,* 314–347.
6. Zhou, H., & Wen, Q. (2014). A new solution of data security accessing for Hadoop based on CP-ABE. In *IEEE 2014*.
7. https://elastic-security.com.
8. Clouder Inc. (2015). *HDFS data at rest encryption*. Retrieved July 12, 2015 from http://www.cloudera.com/content/cloudera/en/documentation/core/latest/topics/cdhsghdfsencryption.html#xd583c10bfdbd326ba--5a52cca-1476e7473cd--7f85.
9. Jayan, A., & Upadhyay, B. R. (2017). RC4 in Hadoop security using mapreduce. In *2017 International Conference on Computational Intelligence in Data Science (ICCIDS).*
10. Jing, F. A. H., Renfa, S. B. L., & Zhuo, T. C. T. (2013). The research of the data security for cloud disk based on the Hadoop framework. In *2013 Fourth International Conference on Intelligent Control and Information Processing (ICICIP).*
11. Jung, Y.-A., & Woo, S.-J. (2015). A study on Hash Chain-based Hadoop security scheme. In *IEEE 2015.*
12. Yu, X., Ning, P., & Vouk, M. A. Enhancing security of Hadoop in a public cloud. In *2015 6th International Conference on Information and Communication Systems (ICICS).*
13. Dean, J., & Ghemawat, S. (2004, December). MapReduce: Simplified data processing on large clusters. In *Proceedings of the 6th Conference on Symposium on Operating Systems Design and Implementaton*, pp. 137–150.
14. HadoopGIS on the FutureGrid.
15. Jam, M. R., Khanli, L. M., & Akbari, M. K. (2014). A survey on security of Hadoop. In *2014 ICCKE*.
16. O'Malley, O., Zhang, K., Radia, S., Marti, R., & Harrell, C. (2009). Hadoop security design. In *Yahoo, Inc., Tech. Rep*.
17. Yuan, M. (2012). Study of security mechanism based on Hadoop. *Information Security and Communications Privacy, 6,* 042.
18. Somu, N., Gangaa, A., & Sriram, V. S. S. (2014, April). Authentication service in Hadoop using one time pad. *Indian Journal of Science and Technology, 7*(4), 56–62.