www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 3 Issue 12 December 2014, Page No. 9507-9509

Review on Big Data Security in Hadoop

Vijaykumar Patil[1], Prof. Venkateshan N[2]

Department of Computer Engineering, University of Pune
SKN'Sinhgad Institute of Technology and Science, Lonavala, Pune, Maharashtra, India.
[1]*vijay.patil.karad@gmail.com*
[2]*venktesann.sknsits@sinhgad.edu*

Abstract: ***These instructions provide you guidelines for preparing papers for International Journal of engineering & computer science Hadoop popularly used for processing large amount of data on its distributed programming framework with Hadoop distributed file system (HDFS), but processing sensitive or personal data on distributed environment demands secure computing. Originally Hodoop was invented without any security model. The encryption and decryption are used before writing and reading data from Hadoop distributed file system (HDFS) respectively. Advanced Encryption Standard (AES) enables protection to data at each cluster, it perform encryption/decryption before read/write respectively. Hodoop running on distributed environment, it uses commodity hardware which is a network model require a strong security mechanism, in addition kerberos used for authentication it offers access control list and audit mechanisms, to ensure the data stored in the Hadoop file system is secure.***

Keywords: Hadoop,DataNode,NameNode,TaskTracker,ASE,HDFS.

1. Introduction

Hadoop was developed from GFS (Google File System) [2, 3] and MapReduce papers published by Google in 2003 and 2004 respectively. It has been popular recently due to its highly scalable distributed programming or computing framework, it enables processing big data for data-intensive applications as well as many analytics. Hadoop is a framework of tools which supports running application on big data and it is implemented in java. It provide MapReduce programming architecture with a Hadoop distributed file system(HDFS), which has massive data processing capability with thousands of commodity hardware's by using simply its map and reduce functions.

Since Hadoop is usually executing in large cluster or may be in a public cloud service. Like Amazon, Google, Yahoo, etc. are such public cloud where multiple users can run their jobs using Elastic MapReduce and cloud storage that is used as HDFS, it is essential to implement the security of user data on such storage or cluster. Hadoop project during its early design stage the simple security mechanisms are employed such as file permissions and access control list [4]. Encryption and decryption is key means for securing Hadoop file system(HDFS), where many DataNodes (or clusters that is originally DataNodes) store file to HDFS, those are transferred while executing MapReduce (user submitted program) job. It is reported that upcoming Hadoop software or version will include encryption and decryption functionality [5].

Internet now generating large amount of data every day, IDC's publish a statistics in 2012 it include the structured data on the internet is about 32% and unstructured is 63%. Also the volume of digital content on internet grows up to more than 2.7ZB in 2012 which is up 48% from 2011 and now rocketing towards more than 8ZB by 2015. Every industry and business organizations are now an important data about different product, production and its market survey which is a big data beneficial for productivity growth.

In commercial data analysis application which is operate on big data the Hadoop becomes de facto platform, in upcoming 5 year, more than 50% of big data applications are executing on Hadoop.

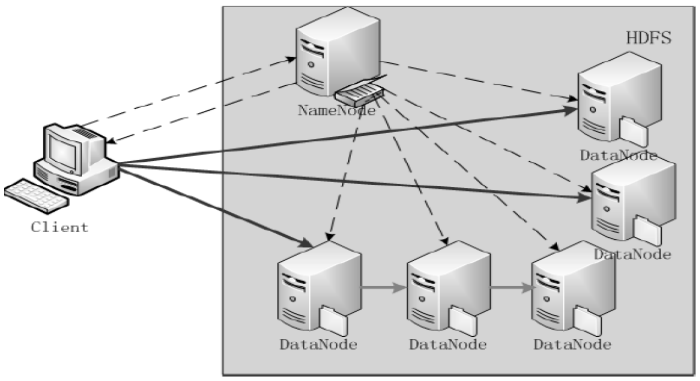

Figure 1: HDFS architecture [11]

Files on Hadoop file system (HDFS) are split into different blocks and replicated with multiple DataNodes to ensure high data availability and durability to failure of execution of parallel application in Hadoop environment. Originally Hadoop clusters have two types of node operating as master-salve or master-worker pattern. NameNode as a master and DataNodes are workers nodes of HDFS. Where data files are actually located in Hadoop is known as DataNode which only leads storage. However NameNode contains information about where the different file blocks are located but it is not persistent, when system starts block may changes one DataNode to another DataNode but it report to NameNode or client who submit the MapReduce job or owner of Data

periodically [11]. The communication is in between DataNode and client NameNode only contains metadata.

2. Security Risks in HDFS

Hadoop uses 'whoami' and 'bash –c groups" utility of Unix for individual user and groups respectively, this is the weak point because which permissions and file quota are for clients. There are three kinds of security violations in HDFS, unauthorized access, unauthorized modification of data and denial of service or resource.

Following are the areas where threat identify in Hadoop

- **Hadoop does enforce authenticate any user or service**: unauthorized users may any HDFS cluster like owner via RPC of HTTP protocol.
- **DataNode can't have any access control mechanism to protect data block** : it is possible to write or modify existing data blocks to DataNode.
- **An attacker can presence as Hadoop service** : For example, code submitted by user register itself on MapReduce cluster as a new TaskTracker
- **Super-user of system does anything without checking**: User who takes control of NameNode is a super-user; it means somebody started the NameNode which have fully access on HDFS data.
- **An executing MapReduce may use the host operating system interfaces**: Some time execution of MapReduce demands access other tasks on the host OS, access local storage for instant Map output, but both executing on the same physical node.

3. Literature Review

Hadoop is originally a distributed system which allows us to store big data and supports for processing it in parallel environment. Many organizations uses big data applications to predict future scope, Hadoop cluster store the sensitive information about such organizations (information like productivity, financial data, customer feedback etc). As result Hadoop cluster require strong authentication and authorization with data protection such as encryption.

The authors Seonyoung Park and Youngseok Lee, they present secure Hadoop architecture by adding encryption and decryption functions in HDFS in " Secure Hadoop with Encrypted HDFS ", J.J. Park et al. (Eds.): GPC 2013, LNCS 7861, pp. 134–141, 2013 Springer-Verlag Berlin Heidelberg. They publish secure HDFS by adding the AES encrypt/decrypt class to CompressionCodec in Hadoop.

The authors Jason Cohen and Dr. Subatra Acharya, they present Trusted Computing Group (TCG), such as the pervasively available Trusted Platform Module (TPM) concerns for achieving data confidentiality and integrity in " Towards a Trusted Hadoop Storage Platform: Design Considerations of an AES Based Encryption Scheme with TPM Rooted Key Protections", IEEE 10th International Conference on Ubiquitous Intelligence & Computing and IEEE 10th International Conference on Autonomic & Trusted Computing in 2013. They publish an encryption scheme for Hadoop utilizing hardware key protections and AES-NI for encryption acceleration.

The authors Hsiao-Ying Lin, Shiuan-Tzuo Shen, Wen-Guey Tzeng and Bao-Shuh P. Lin, they present the data confidentiality issue by integrating hybrid encryption schemes in the Hadoop distributed file system (HDFS) in " Toward Data Confidentiality via Integrating Hybrid

Encryption Schemes and Hadoop Distributed File System", 26th IEEE International Conference on Advanced Information Networking and Applications, 2012 Springer-Verlag Berlin Heidelberg. They publish two integrations, HDFS-RSA and HDFS-Pairing, as extensions of HDFS.

The authors Thanh Cuong Nguyen, Wenfeng Shen, Jiwei Jiang and Weimin Xu, they present the scheme to encrypt user's data locally before transferring to HDFS if he requires high privacy" A Novel Data Encryption in HDFS ", IEEE International Conference on Green Computing and Communications and IEEE Internet of Things and IEEE Cyber, Physical and Social Computing, 2013. They publish novel method to encrypt file while being uploaded.

The authors Songchang Jin, Shuqiang Yang, Xiang Zhu, and Hong Yin, they present the latest cryptography—fully homomorphic encryption technology and authentication agent technology in " Design of a Trusted File System Based on Hadoop ", Y. Yuan, X. Wu, and Y. Lu (Eds.): ISCTCS 2012, CCIS 320, pp. 673–680, 2013. They publish homomorphic encryption and authentication agent technology for protecting HDFS.

The authors Monika Kumari and Dr.Sanjay Tyagi, they present three stage security model is presented for Hadoop Environment in " A Three Layered Security Model for Data Management in Hadoop Environment", International Journal of Advanced Research in Computer Science and Software Engineering Volume 4, Issue 6, June 2014. They publish secure file management and distribution over the secure Hadoop environment.

4. Secure Hadoop

Hadoop architecture consists of a master and all others are slaves. Master contains NameNode that manages metadata and access control to file system for mapping, DataNode and block of file, slaves are DataNode which store data. The HDFS contains data in block of fixed size, by default block size is 64 MB. Each block is replicated three times in different DataNode, even after processing or every time Hadoop maintains replication factor three. Hadoop provide MapReduce programming model which split job into multiple tasks (map or reduce) to process more than one HDFS data blocks in parallel. HDFS supports a write-once-read-many model.

Secure Hadoop encrypt every file before written in HDFS. It is reported that every DataNode or slave is a commodity server which perform encryption or decryption at local site using its CPUs [1]. Advanced Encryption Standard (AES) is most popular algorithm that support block cipher, hence it is suitable for HDFS blocks. AES available with 128-bit AES, 192-bit AES and 256-bit AES, 128-bit AES is used most of times because of its simplicity. There are different modes of operations of AES: ECB, OFB, CTR, XTS and CBC. It is reported that AES: ECB is good choice of encryption or decryption algorithm because its concurrently performed a computation in a distributed environment [1].

4.1 Encryption in HDFS

Figure 2 shows operation that save every block in to HDFS, client split each file in to fixed size block and encrypts it before upload to Hadoop file system. It is reported that encryption and decryption can be implemented simply by using Java class [1]. Clients, itself perform encryption using AES algorithm on the CPU and transfer encrypted block to HDFS

(DataNode). Then receiver DataNode (First DataNode where block store) replicate block in to two other DataNodes.

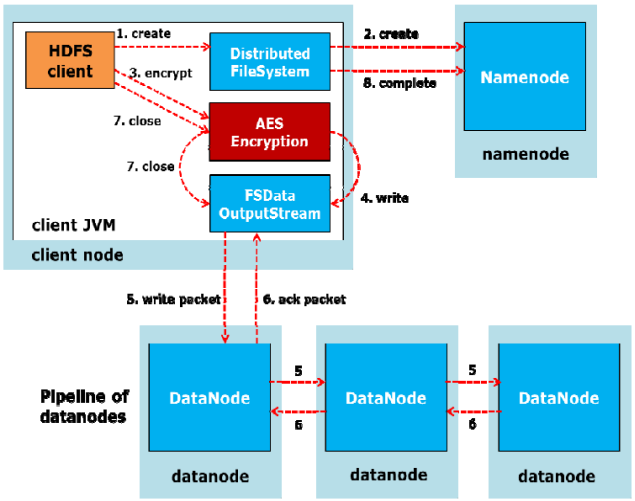

Figure 2: Writing a file by adding an encryption step[1]

4.2 DECRYPTION IN HDFS

Data blocks are written by client to DataNode sequentially, but during execution of MapReduce job multiple blocks are read (decrypted) parallel at TaskTracker. Figure 3 shows that MapTask read and encrypt data blocks at TaskTracker using AES encryption method. It is reported that multiple MapTasks are executing in Hadoop at worker sites. HDFS supports write-once-read-many model, it is reported that concurrent decryption of HDFS block well suitable for many MapReduce jobs [1].

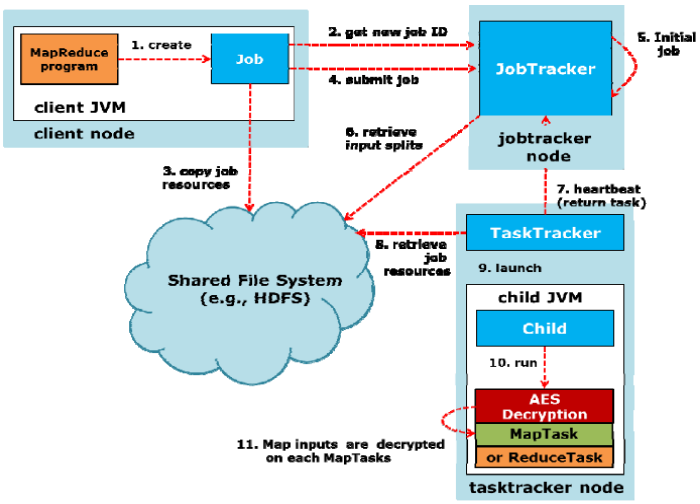

Figure 3: A MapReduce job that read an encrypted file[1]

5. Future Scope

Big data contains sensitive and private information, in order to protect this big volume that stored at different commodity hardware, necessary to implement authentication to verify user or system identity. Authorization is useful for providing access control privileges to user or system; also the ACL's are helps for file permission. OAuth 2.0 is good choice for both authentication and Authorization. Additionally, audit trails used for tracking each user activity. OAuth 2.0 token powerful mechanism that support AES to provide data confidentiality and integrity among different user

6. Conclusion

In the era of Big Data, where data is collected from different sources, security is a measure issue, as there no any fixed source of data and not any kind of security mechanism. Hadoop adopted by various industries to process such data, demands strong security solution. Thus authentication, authorization and encryption or decryption methods are much helpful to secure Hadoop file system.

References

[1] Seonyoung Park and Youngseok Lee "Secure Hadoop with Encrypted HDFS"

[2] Dean, J., Ghemawat, S.: MapReduce: Simplified Data Processing on Large Cluster. In:OSDI (2004)

[3] Ghemawat, S., Gobioff, H., Leung, S.: The Google File System. In: ACM Symposium onOperating Systems Principles (October 2003)

[4] O'Malley, O., Zhang, K., Radia, S., Marti, R., Harrell, C.: Hadoop Security Design,Technical Report (October 2009)

[5] White, T.: Hadoop: The Definitive Guide, 1st edn. O'Reilly Media (2009)

[6] Hadoop, http://hadoop.apache.org/

[7] Jason Cohen and Dr. Subatra Acharya "Towards a Trusted Hadoop Storage Platform:Design Considerations of an AES Based Encryption Scheme with TPM Rooted KeyProtections" (2013)

[8] Lin, H., Seh, S., Tzeng, W., Lin, B.P. " Toward Data Confidentiality via Integrating sfsfHybrid Encryption Schemes and Hadoop Distributed FileSystem" (2012)

[9] Thanh Cuong Nguyen, Wenfeng Shen, Jiwei Jiang and Weimin Xu "A Novel Data Encryption in HDFS" (2013)

[10] Devaraj Das, Owen O'Malley, Sanjay Radia and Kan Zhang "Adding Security to Apache Hadoop"

[11] Songchang Jin, Shuqiang Yang, Xiang Zhu, and Hong Yin "Design of a Trusted File System Based on Hadoop " 2013

[12] Advanced Encryption Standard, http://en.wikipedia.org/wiki/Advanced_Encryption_Standard

[13] Sharma Y. ; Kumar S. and Pai R.M; "Formal Verification of OAuth 2.0 Using Alloy Framework "

[14] Ke Liu and Beijing Univ "OAuth Based Authentication and Authorization in Open Telco API "