IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 10, Issue 1, October 2021
Impact Factor: 5.731

Review Paper on Security of Hadoop

Kalyani Jha
Student MSc, Department of Information Technology
Sir Sitaram and Lady Shantabai Patkar College of Arts and Science, Mumbai, India
kalyanijha50@gmail.com

Abstract: *Trusted computing and security of services is one of the most challenging topics today and is the cloud computing's core technology that is currently the focus of international IT universe. Hadoop, as an open-source cloud computing and big data framework, is increasingly used in the business world, while the weakness of security mechanism now becomes one of the main problems obstructing its development. This paper first describes the Hadoop project and its present security mechanisms, then analyzes the security problems and risks of it, pondering some methods to enhance its trust and security and finally based on previous descriptions, concludes Hadoop's security challenges.*

Keywords: Security; Trust; Hadoop; Big Data; MapReduce; Cloud Computing

I. INTRODUCTION

In the current information age, the requirement of data is increasing day by day. The data generated from different sources is in terabytes per day, which is called as big-data. Big-data is not just big in size, but big data have data of different variations, different sizes and at different speed. This big-data is used for many applications and business-related services like business intelligence. To store and process this large amount of data we need an efficient and fault tolerant system. Google developed a file system called as Google file system to handle big data. Hadoop is based on the Google's file system. Hadoop is open-source software framework to store and process this big data efficiently. It is designed in java language. HDFS (Hadoop Distributed File System) and Map Reduce are the two components of Hadoop.[1]

Today, data explosion is a reality of digital universe and the amount of data extremely increases even in every second. IDC's latest statistics show that rate of structured data in the Internet now have been grown about 32%, and unstructured data about 63%. To 2012, the unstructured data will occupy more than 75% proportion of the entire amount of data in the Internet. [2]

The volume of digital content of the world grows to 8ZB by 2015.One common programming model to handle and process this extreme amount of Big Data is MapReduce. Apache Hadoop is an opensource software framework and well-known implementation of MapReduce model that supports data-intensive distributed applications. Hadoop Distributed File System (HDFS) is a distributed, scalable, and portable file system written in Java for the Hadoop framework, and actually it is cloud storage the most widely used tool. In fact, in the next 5 years, 50 percent of Big Data projects expect to be run on Hadoop. Financial organizations using Hadoop started to store their confidential sensitive data on Hadoop clusters. So, a need for a strong authentication and authorization mechanism to protect the sensitive data is observed and also there is a need for a highly secure authentication system to restrict the access to the confidential business data that are processed and stored in an open framework like Hadoop.[2]

However, whether the user is assured that put the private data to various clouds is the key to widely promotion of cloud storage. Initially, Hadoop had no security framework and it considers that the entire cluster, user and the environment were trusted. Even though it had some authorization controls like file access permissions, a malicious user can easily impersonate a trusted used as the authentication were on the basis of Password. Later on, Hadoop cluster moved on to private networks, where the users have equal rights to access the data stored in the cluster.

Equal access to all users gives the malicious user the possibility of firstly, read or modify the data in the other's cluster and secondly, suppress or kill the other job to execute his job earlier than the other to complete job from a malicious user because the data node does not enforce access control policies.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 10, Issue 1, October 2021
Impact Factor: 5.731

II. HADOOP SYSTEM

The current Hadoop framework does not support two important features first is encryption of storing HDFS blocks and computation on such encrypted data which is a fundamental solution for secure Hadoop, and second is if same data is occurred then what should be the processing strategy. To overcome these two problems, we need a principled way for the encryption process, and to minimize the time of file encryption and job execution (file decryption) and compare duplicate input data to avoid processing of samedata multiple times. Input to proposed system is multiple numbers of files; the system will first encrypt files and then load at HDFS, then execute the job on data at HDFS on user request. At the time of job execution; it needs to perform decryption.[1]

File on HDFS splits into multiple blocks and replicated into multiple Data Nodes to ensure high data availability and durability to avoid failure of execution for parallel application in Hadoop environment. Originally Hadoop clusters have two types of nodes i.e., master-slave. Name Node as a master and Data Nodes are workers nodes of HDFS. Data files which are located in Hadoop are stored in Data Node which only stores data. However, Name Node contains information about where the different file blocks are located but it is not persistent, when system starts block may changes one Data Node to another Data Node but it reports toName Node or client who submit the Map Reduce job or owner of Data periodically. The communication is in between Data Node and client Name Node only contains metadata. [1]

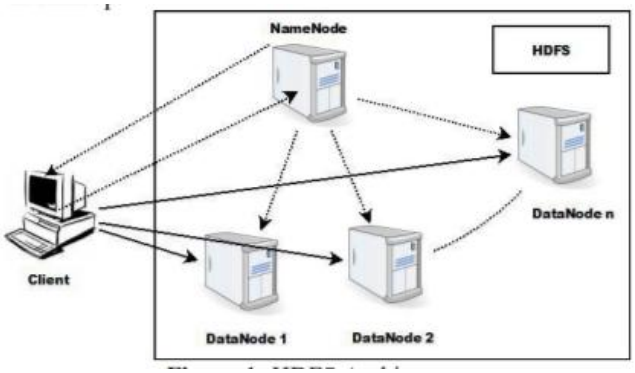

Figure 1: HDFS Architecture

The Apache Hadoop project develops open-source software for reliable, scalable, distributed computing. It is a framework that by the use of plain programming models, permits for the distributed parallel processing of big data sets in the size of petabytes and exabytes across clusters of computers so that a cluster of Hadoops can easily scale out and also scale up from single servers to thousands of machines which each of them offer local computation and storage. Many companies like amazon, Facebook, yahoo, etc. store and process their data on Hadoop that proves its popularity and robustness. The library itself is designed to detect and handle failures at the application layer, rather than rely on hardware to deliver high-availability, so delivering a highly-available service on the upside of a cluster of computers, each of which may be prone to failures.

However, enterprises want to protect sensitive data, while the weakness of security mechanism now becomes one of the main problems obstructing Hadoop's development and use and because of the lack of a valid user authentication and data security defense measures, Hadoop is now facing many security problems in the data storage.

III. 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 [10] 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

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 10, Issue 1, October 2021
Impact Factor: 5.731

service has a Kerberos "principal" and credentials are by Service: keytab 1 s and User: password which RPC 2 Encryption should be enabled.

3.1. Apache Knox Gateway

The Apache Knox Gateway is a system that provides a single point of authentication and access for Apache Hadoop services. It accesses over HTTP/HTTPs to Hadoop Cluster and provides the following features.

- Single REST API Access Point
- Centralized authentication, authorization and audit for Hadoop REST/HTTP services
- LDAP 4 /AD 5 Authentication, Service Authorization and Audit
- Eliminates SSH edge node risks
- Hides Network Topology.

3.2. Authentication

Authentication means to identify who you are. Providers with the role of authentication are responsible for collecting credentials presented by the API consumer, validating them and communicating the successful or failed authentication to the client or the rest of the provider chain. By this primer security, untrusted users do not have access to the cluster network and trusted network, everyone is good citizen. Your identity is determined by client host. For strong authentication, Hadoop uses:

- Kerberos
- LDAP, Active Directory
- LDAP, AD integrated with Kerberos, establishing a single point of truth
- Single point of truth

Kerberos is a computer network authentication protocol which works on the basis of "tickets" to allow nodes communicating over a non-secure network to prove their identity to one another in a secure manner.

In the development of HDFS cluster, it places the trusted server authentication key in each node of the cluster to achieve the reliability of the Hadoop cluster node communication, which can effectively prevent non-trusted machines posing as internal nodes registered to the NameNode and then process data on HDFS. This mechanism is used throughout the cluster. So from storage perspective, Kerberos can guarantee the credibility of the nodes in HDFS cluster.

Kerberos can be connected to corporate LDAP environments to centrally provision user information. Hadoop also provides perimeter authentication through Apache Knox for REST APIs and Web services. However, Kerberos is ineffective against Password guessing attacks and does not provide multipart authentication.

3.3. Authorization

Authorizationis the process of ensuring that users have access only to data as per corporate policies. Hadoop already provides fine-grained authorization via file permissions in HDFS, resource-level access control for YARN and MapReduce, and coarser-grained access control at a service level.

The authorization role is used by providers that make access decisions for the requested resources based on the effective user identity context. This identity context is determined by the authentication provider and the identity assertion provider mapping rules. Evaluation of the identity contexts user and group principals against a set of access policies is done by the authorization provider in order to determine whether access should be granted to the effective user for the requested resource.

Out of the box, the Knox Gateway provides an ACL based authorization provider that evaluates rules that comprise of username, groups and ip addresses. These ACLs are bound to and protect resources at the service level. That is, they protect access to the Hadoop services themselves based on user, group and remote ip address. To provide a common authorization framework for the Hadoop platform, providing security administrators with a single administrative console to manage all the authorization policies for Hadoop components is the goal of Hadoop's developers.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 10, Issue 1, October 2021
Impact Factor: 5.731

3.4. OS Security and Data Protection

Data protection involves protecting data at rest and in motion, including encryption and masking. Encryption provides an added layer of security by protecting data when it is transferred and when it is stored (at rest), while masking capabilities enable security administrators to desensitize PII for display or temporary storage. In Hadoop it will be continued to leverage the existing capabilities for encrypting data in flight, while bringing forward partner solutions for encrypting data at rest, data discovery, and data masking.

IV. Security Threats In Hadoop

Hadoop does not follow any classic interaction model as the file system is partitioned and the data resides in clusters at different points. One of the two situations can happen: job runs on another node different from the node where the user is authenticated or different set of jobs can run on a same node. The areas of security breach in Hadoop are:

- Unauthorized user can access the HDFS file
- Unauthorized user can read/write the data block
- Unauthorized user can submit a job, change the priority, or delete the job in the queue.
- A running task can access the data of other task through operating system interfaces

Some of the possible solutions can be:

- Access control at the file system level
- Access control checks at the beginning of read and write
- Secure way of user authentication

Authorization is the process of specifying the access right to the resources that the user can access. Without properauthentication service one cannot assure proper authorization. Password authentication is ineffective against:

- Replay attack - Invader copies the stream of communications in-between two parties and reproduces the same to one or more parties.
- Stolen verifier attack - Stolen verifier attack occur when the invader snips the Password verifier from the server and makes himself as an legitimate user.

V. Apache Sentry

There is an option to secure Hadoop cluster with Apache Sentry. Sentry is a highly modular system to provide fine grained role-based authorization to both data and metadata stored on an Apache Hadoop cluster. It provides authorization required to provide precise levels of access to the right users and applications. Sentry's key benefits include store sensitive data in Hadoop, extend Hadoop to more users, create new use cases for Hadoop and comply with regulations. Also, its key capabilities of it are Fine-Grained authorization, Role-Based authorization, Multi-Tenant administration, to separate policies for each database/schema and ability to be maintained by separate admins. Sentry have been proposed and launched by Cloudera Company. [6,7]

VI. Fully Homomorphic Encryption

This research (Jam, Akbari & Khanli, 2014) proposed a design of trusted file system by combining the authentication agent technology with the cryptography fully homomorphic encryption technology. This is used for Hadoop which provides reliability and security from data, hardware, users and operations. This enables the user to prevent data breach along with enhanced efficiency of the application which is possible due to the encrypted data in the homomorphic encryption technology. Authentic agent technology also provides a range of techniques which are an integration of different mechanisms such as privilege separation and security audit that provides security of data in Hadoop system. [5]

Fully homomorphic encryption allows multiple users to work on encrypted data in an encrypted form with any operation, but yields the same results as if the data had been unlocked. So, it can be used to encrypt the data for users, and then, the encrypted data can be uploaded to HDFS without worrying that data be stolen when transferring on the network to HDFS. After data processing with MapReduce, the result is still encrypted and safely stored on HDFS.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 10, Issue 1, October 2021
Impact Factor: 5.731

VII. Authentication Using One Time Pad

A novel and a simple authentication model using one time pad algorithm is proposed that removes thecommunication of passwords between the servers. This model tends to enhance the security in Hadoop environment.

The proposed approach provides authentication service by using one time pad and symmetric cipher cryptographic technique. This approach uses two-server model, with a Registration Server and a Back end Server. The whole process of authentication consists of two parts:

- Registration Process
- Authentication Process

During the registration process, the user enters his Username and Password. The Password is encrypted (Cipher Text 1) using one-time pad algorithm. Cipher Text 1 is again encrypted using mod 26 operations (Cipher Text 2) and stored in the Registration Server. Again, encrypt the onetime pad key using the Password which results in (Cipher Text 3) using symmetric cipher technique. Cipher Text 3 will be sent to the Backend Server to be stored along with the Username.

Next during the authentication process, after receiving the Username from the user, the Registration Server sends the Username to the user. The Backend Server sends the corresponding Cipher (Cipher Text 3) to the User via Registration Server. The user deciphers it using his Password and returns the key to Registration Server.

Registration Server decrypts Cipher Text 1 with the key returned by the User. Again, encrypts the Password with same key and send the Cipher (Cipher Text 4) to the Backend Server.

The Backend server compares Cipher Text 4 with Cipher Text 3. If it matches, sends the Username to the Registration Server. The Registration Server compares the Username with the Username entered by the user. If it matches, the user is authenticated. The random is valid only for one session. Once the user logs out, a new random key replaces the old one.

VIII. Triple Encryption Scheme for Hadoop-Based Data

Cloud computing has been flourishing in past years because of its ability to provide users with on-demand, flexible, reliable, and low-cost services. With more and more cloud applications being available, data security protection becomes an important issue to the cloud. In order to ensure data security in cloud data storage, a novel triple encryption scheme is proposed in this paper, which combines HDFS files encryption using DEA and the data key encryption with RSA, and then encrypts the user's RSA private key using IDEA.

A novel triple encryption scheme is proposed and implemented, which combines HDFS files encryption using DEA (Data Encryption Algorithm) and the data key encryption with RSA, and then encrypts the user's RSA private key using IDEA (International Data Encryption Algorithm).

In the triple encryption scheme, HDFS files are encrypted by using the hybrid encryption based on DES and RSA, and the user's RSA private key is encrypted using IDEA. The triple encryption scheme is implemented and integrated in Hadoop-based cloud data storage.

Principle of Data Hybrid Encryption is that HDFS files are encrypted using a hybrid encryption method, a HDFS file is symmetrically encrypted by a unique 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. Hybrid encryption is a compromising choice against the two forms of encryption above. Hybrid encryption uses DES algorithm to encrypt files and get the Data key, and then uses RSA algorithm to encrypt the Data key. User keeps the private key in order to decrypt the Data key.

They have planned to achieve the parallel processing of the encryption and decryption using MapReduce, in order to improve the performance of data encryption and decryption.

IX. Conclusion

In this paper, we reviewed security of Apache Hadoop platform including its present security situation, threats and some methods enhancing its security level. We can claim that Hadoop has strong security at the file system level, but it

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 10, Issue 1, October 2021
Impact Factor: 5.731

lacks the granular support needed to completely secure access to data by users and Business Intelligence applications. This problem forces organizations in industries for which security is paramount (such as financial services, healthcare, and government) to choose either leave data unprotected or lock out users entirely. Mostly, the preferred choice is the latter, severely inhibiting access to data in Hadoop. Although to solve the problem of secure access to data, Apache Sentry has been newly proposed and it promises to be successful, due to overcome these difficulties and make Hadoop secure for enterprises, actually new methods are needed to be proposed.

REFERENCES

[1]. Swapnali A. Salunkhe , Amol B. Rajmane, Department of Computer Science and Engineering, Ashokrao Mane Group of Institutions, Vathar, Maharashtra, India;https://www.ijsr.net/archive/v6i7/ART20175080.pdf

[2]. Masoumeh RezaeiJam, Leili Mohammad Khanli, Mohammad Kazem Akbari, A Survey on Security of Hadoop; https://ieeexplore.ieee.org/document/6993455

[3]. https://securosis.com/assets/library/reports/Securing_Hadoop_Final_V2.pdf

[4]. https://www.researchgate.net/publication/323790230_An_approach_for_big_data_security_based_on_Hadoop_distributed_file_system

[5]. Gayatri Kapil, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia, https://peerj.com/articles/cs-259/

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

[7]. (2014). Security for Hadoop. Available: http://www.cloudera.com/content/cloudera/en/solutions/enterprise solutions/security-for-hadoop.html