Volume 7, Issue 1, January 2017 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Detailed Review on Securing Hadoop HDFS using Kerberos Protocol

Aarti Kalsi
Dept. of Computer Science and Engineering
Sri Sai College of Engineering and Technology
Badhani, Punjab, India

Anshu Chopra
Dept. of Computer Science and Engineering
Sri Sai College of Engineering and Technology
Badhani, Punjab, India

DOI: 10.23956/ijarcsse/V6I8/0110

Abstract: ***Data driven ventures perceive that the cutting edge of information administration is a solitary, focal framework that stores all information in any structure or sum and furnishes business clients with a wide cluster of apparatuses and applications for working with this information. This advancement in information administration is the undertaking information centre point, and Apache Hadoop is at its centre. Yet as associations expanding store information in this framework, a developing segment is business-delicate and subject to regulations and administration controls. This data requires that Hadoop give solid abilities and measures to guarantee security and consistence. This paper expects to investigate and outline the basic security methodologies, devices, and practices found inside Hadoop everywhere, and how layered approach is particularly situated to give, reinforce, and deal with the information security required for an endeavour information centre point.***

Keyword: Hadoop, HDFS, Kerberos, MapReduce

I. INTRODUCTION

The amount of data is increasing day by day. This data is present in terabytes or petabytes. For handling this huge amount of data is not simple task. For analyzing this data we use hadoop technology. With the help of this technology we use HDFS and MapReduce paradigms for handling the computational requirements. After analyzed the data, there is a requirement of security. Without secure connections, no one use the technology because users need secure account for privacy purpose. So i analyzed some hadoop powers with respect to security aspects. In today's world, hadoop gets a lot of attention and it is a buzz for all. Even all the technologies are somewhere based on it, but still it's a grey area for many users. So we discuss about the Hadoop technology and its security features.

Hadoop works with volumes of data[1]. It is an open source platform which is designed for store the data. it's a huge quantity of data which is stored by it and also processed by it. Nowadays quantity of data is increasing exponentially. Servers of data can be increased in a single day so it is not possible to handle or buy the servers on daily basis. But hadoop is one that can not only store the data but processes it. It is easily scaled up and down for the data and also the shared memory or we can say disk space does not shared by any type of commodity servers.

Hadoop works with two types of nodes one is worker node and other is slave node. Both the nodes can work simultaneously. So there is not a single minute that hadoop does not work. Generally the huge amount of work is splited into number of tasks and these tasks can be completed in different places with the help of distributing system.

There are four important terms of hadoop i.e. **open source software, framework, massive storage and processing power.**

The term **open source** means that we can download the software free which is maintained by the developers at globally.

Framework is nothing but provides a platform at which we can develop the software and applications. Hadoop generally splits the huge data into sub blocks and those **splitted** data are stored on different clusters of servers.

The **processing power** of hadoop is fast because a number of systems are connected to each other so that work easily processed.

There are generally four important parts of hadoop system [2]. They are as follows:

1. Hadoop common
2. Hadoop Distributed File System (HDFS)
3. MapReduce
4. YARN

HDFS

It is one by which huge amount of data is stored into the system [3]. It was developed with the help of distributed system design. The commodity hardware is the one by which it can be run. The important point for HDFS is that it is highly fault tolerant and the cost for designing the system is not much more.

To store and access the data is an easy task with it. This huge data is stored in various machines. The way of storing the data in HDFS is in redundant fashion by which it becomes fault tolerant. Hadoop has the power to create interface by which interacting with HDFS is possible.

MapReduce

This is the technique based on distributed computing. As the name MapReduce, consists of two words Map and Reduce. The working of map is to take the data sets and convert this data set into another by splitting the data into number of pairs. Reduce means to reduce the data sets into logical output form without changing the meaning of the data set.

These are the points, why we use hadoop technology. They are as follows:

- The low cost is the important thing that makes it useful worldwide by which storing and processing the data in an efficient way. We can store as much as data in the commodity hardware because we can easily buy a number of servers to store the data.
- One of the important uses of the hadoop that a large amount of initial data goes for analytical purpose or we can say that transferring of data in a warehouse system. Nowadays all the organizations want to process the historical data into the knowledge data.
- The other important point is data lake that means it stores the data without any special commands or queries that we use the oracle or sql world. The platform that has ETL power use the data management system by which refined results are generated according to our requirements.
- Hadoop performs like a sandbox. Generally the hadoop is used for huge amount of data which shapes our data into knowledgeable data with the help of various algorithms and analytical system. it has the power to create new opportunities in this competitive world.
- The most important task of hadoop is the web based recommendation system. People use facebook, Linkedln, twitter and for search a single web page it stores a huge amount of data based on real time.

The first question arises the how we can insert our data into hadoop and how the security performs for it.

- The firsts step to load the files to the system with the help of some java commands. After inserting, it is the responsibility of HDFS to make the multiple data blocks and distributes those blocks to different systems for processing parallel.
- Suppose you have a large number of files then we use the script called shell script. For the multiple computational operations, this script process number of PUT commands. So if you do this then there is no need to use MapReduce commands.
- For scanning the directory, we have to create cron job for the new files and then transfer these files into the HDFS. This technique is important for receiving emails at regular intervals.
- Now these files put into the HDFS system. Now we have to use sqoop for import the information from the raw data.
- We can also use FLUME by which we can transfer the load from logs into hadoop.

What is security

Security is nothing but to protect the information from an unauthorized access or any disaster and only the authorized persons should allow the access their data [4]. It is responsibility of the technology or system to protect the data from publication or unauthorized access.

Hadoop security:

Hadoop does not follow the strong access control mechanism and it does not authenticate the users with any special command. The basic cons of hadoop technology are that any type of user can communicate with the datanode without any authorization of namenode and can easily get the details of blocks of data.

There are some challenges of hadoop security. They are as follows:

Authentication: The first and foremost challenge is how hadoop can ensure that they are the authorized persons and giving them the proper accessing power of account?

Access Control: the second challenge is how hadoop can ensure that users are accessing only their information under the policies?

Auditing: The third challenge is how can you say that the data which is recorded by the hadoop is used only then when worst condition happens?

Data protection: The last but not the least challenge is that is the users data send via encryption method or not or a normal sending procedure?

Network level: The challenges that can be categorized under a network level deal with network Protocols and network security, such as distributed nodes, distributed data, Internodes Communication.

Data level: The challenges that can be categorized under data level deals with data integrity and availability such as data protection and distributed data [6].

Generic types: The challenges that can be categorized under general level are traditional security tools, and use of different technologies

As we know that we can insert the raw data into the hadoop that means it does not authenticates the users by which security is compromised.

1. Any user can use the clusters of HDFS or MapReduce like any other user. This type of right can be compromised in corporative sectors. For example, on HDFS, the permission of the file checking can be easily outwitted.
2. An unauthorized user can easily hack the hadoop services. It means the code which is used by a user can be registered as Task Tracer.

The data blocks in hadoop can easily access by any other person because there is not any accessing control algorithm on the nodes or data nodes.

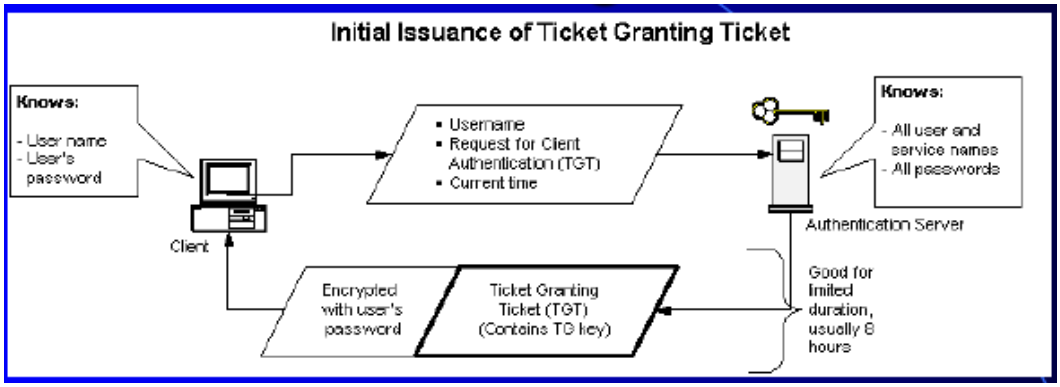
Now the question arises that what should be the requirements for hadoop that no one can access the others data. Some are as follows:

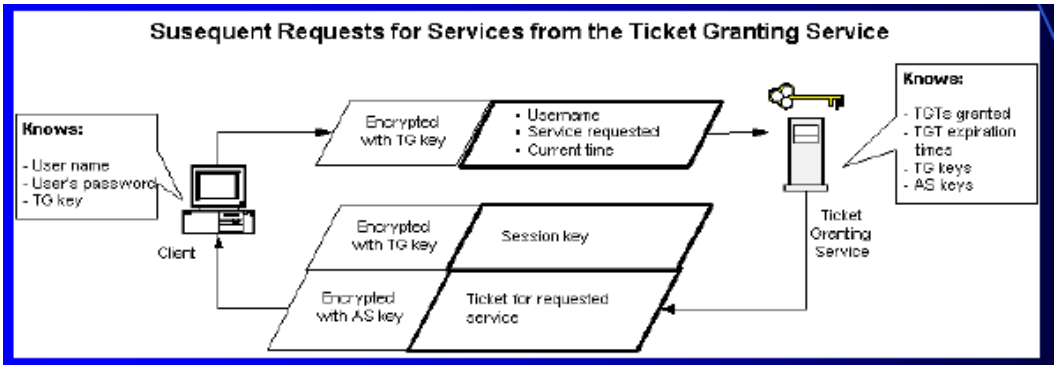
a. Users should be allowed access rights only for their HDFS documents.
b. They should only the right to update their MapReducing jobs.
c. Developers should create an appropriate algorithm for the unauthorized access of datanodes, task tracers.
d. Every user should have their Kerberos credentials or we can say the credentials should be transparent for the users.
e. The last but not the least that, there should be no more decrement in the performance in the GridMix.

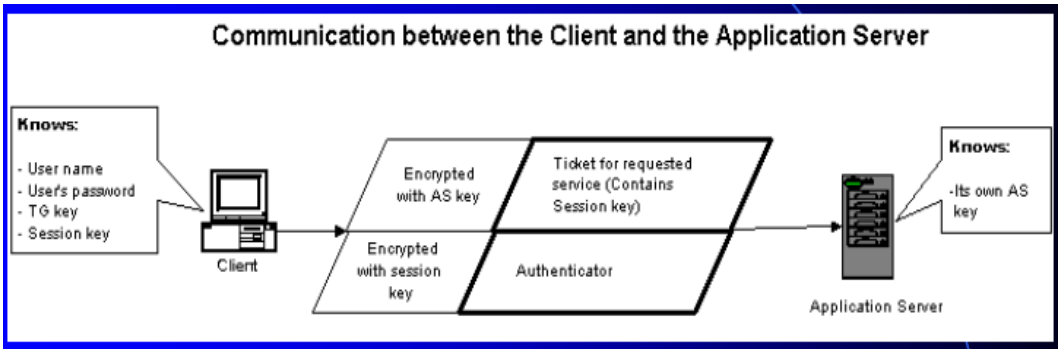
We can design the system for authentication purpose with the help of Kerberos and SSL techniques [4]. It is because for the following reasons:

1. Kerberos has the capability for the symmetric key operations which are generally faster than the public key operations which is used by SSL [5].
2. There should be a simple centrally management policy for deleting the user via key distribution centre and a revocation certification should be transferred to system or servers that should be based on SSL.

There is an example how Kerberos works with Ticket Granting Tickets:







This is the process of Kerberos of Ticket Granting Tickets:

1. For accessing a server of another system. Before processing your request by Kerberos, You must know the particular server that contains a Kerberos TGT or Tickets.
2. For getting ticket, one will have to send a request of authentication from the Authentication Server (AS). Then this server creates a key called "session key" or "encryption Key". This key is based on your password which one get during entering the user name. TGT will be according to the session key.
3. Now the Ticket Granting Server will process your ticket which will be sent by you. The TGS may be physically the same server as the Authentication Server, but it's now performing a different service. The TGS returns the ticket that can be sent to the server for the requested service.
4. The service either rejects the ticket or accepts it and performs the service.
5. Because the ticket you received from the TGS is time-stamped, it allows you to make additional requests using the same ticket within a certain time period (typically, eight hours) without having to be re-authenticated. Making the ticket valid for a limited time period makes it less likely that someone else will be able to use it later.

II. CONCLUSION

As the data is increasing exponentially day by day, the technology is also updating in a same way to handle those huge data. The hadoop technology is one of them. If the technology becomes updated then it needs also security because generated data is valuable for those who create it. Security is nothing but to escape data from a person who want to get it from an unethical way. There are number of technology by which handling of data is possible in an efficient way. The hadoop technology splits the data and then managed it via different technology like HDFS, MapReduce or YARN. It is also possible that a technology creates which is somewhat is good than Hadoop but important thing is that security of data should not be compromise in anyway. There some Hadoop security guides like CDH4 that prevents the users data from malicious which is somewhat like Kerberos.

REFERENCES

[1] Bhosale, Harshawardhan S., and Devendra P. Gadekar. "AREVIEW PAPER ON BIG DATA AND HADOOP." *International Journal of Scientific and Research Publications* (2014): 756.

[2] Smith, Kevin T. "Big Data Security: The Evolution of Hadoop's Security Model." (2013). Smith, Kevin T. "Big Data Security: The Evolution of Hadoop's Security Model." (2013).

[3] Hamlen, Kevin, et al. "Security issues for cloud computing." *Optimizing Information Security and Advancing Privacy Assurance: New Technologies: New Technologies* 150 (2012).

[4] Jin, Songchang, et al. "Design of a trusted file system based on hadoop."*International Conference on Trustworthy Computing and Services*. Springer Berlin Heidelberg, 2012.

[5] http://www.sas.com/en_my/insights/big-data/hadoop.html

[6] Shvachko, Konstantin, et al. "The hadoop distributed file system." *2010 IEEE 26th symposium on mass storage systems and technologies (MSST)*. IEEE, 2010.

[7] Inukollu, Venkata Narasimha, Sailaja Arsi, and Srinivasa Rao Ravuri. "Security issues associated with big data in cloud computing." *International Journal of Network Security & Its Applications* 6.3 (2014): 45.

