A Comparative Study of Various Security Threats and Solutions for the Security of Hadoop Framework In Terms of Authentication and Authorization

Rahul Shahane, Proddutur Shruthi, B.R.V.V.N.L. Viswanadh, Dirisala Abhilash

Abstract: ***In earlier days traditional relational database systems were used.This system cannot handle unstructured data generated by social media or some organization. To process this data Hadoop is used.Big data is the collection of large set data which hold the secured information which can be user's medical records, security data, personal information etc.While we come to Hadoop it is used to store, manage and distribute the data among various nodes.In this we mainly focus on the security threats which arises in big data on the storage purpose and the techniques by which it can be resolved. This paper deals with the problems in security involved with big data in context with the Hadoop environment and the various solution techniques and technologies involving in securing the big data Hadoop.***

Index Terms: Big data, Clusters, HDFS, unauthorized client, unauthorized user.

I. INTRODUCTION

As the word spells big means large we can define big data as large, volume and produce unstructured data from different sources such as sensors, social media etc [1]. At present days the use of data is increasing rapidly due to use of data everywhere such as online shopping, twitter, Facebook, WhatsApp, mobiles [1]. The volumes of big Data haveincreased manifold, as per studies conducted in the year 2012, in a single data set few scores of terabyte data is stored which hasgradually increased to many petabytes today. This large data is used for commercial purpose by enterprise to increase their profits. So, there is need to secure such a large amount of data.

II. BIG DATA

Big data originates from many sources including cell phone gps signal, purchases made online, posts on social media. [8] Hadoop is used for processing vast amounts of data at a swift pace. Hadoop distributed file system supports faster data transfer allowing for uninterrupted system operation at the point of node failure [2].

Properties of big data are:

1. **Volume**: On splitting the term big data, big defines the size of the data. The size of data is known as volume. The size of data can increase from petabytes to zettabytes in near future.
2. **Velocity**: Meaning for velocity is speed. Velocity in Big data means speed of data coming from different various sources.
3. **Variety**: Meaning for the word variety means different. In Big data a scale for measuring the richness of the data representation is variety.
4. **Veracity**: It refers to the noise and the abnormalities of the data. The big data team should keep away the dirty data from gathering around in your systems. It refers to the data quality and data value.

Data must be processed with analytics and algorithms to reveal a significant information. For example to run a company one must think about the visible and invisible issues with multiple components.

III. HADOOP

Hadoop is an open source software frame work for stocking data and managing applications on clusters of hardware. It provides an enormous storage of any kind of data and the ability to handle infinity jobs or tasks. It is a platform which has a huge volume of data which we can say as big data. It is a java-based framework which can store a huge capacity of data sets in a cluster. This frame-work will allow to process in all the nodes and run on a parallel cluster.

IV. HADOOP ECOSYSTEM

Hadoop ecosystem mention about different components of the apache Hadoop software library in addition to the tools and accessories given by the apache software foundation for the software projects.it is a java-based framework that is especially accepted for handing and scrutinizing the huge sets of data.They are different parts in Hadoop Ecosystem. These components make Hadoop so powerful. They are classified as Distributed file system, analytic platform, data storage [8]. The present Hadoop network is made up of Map-Reduce, Hadoop kernel, the Hadoop distributed file system (HDFS) and many related components such as Oozie, apache Hive, pig, Zookeeper and HBase and these components will be explained [7,8]

A. HDFS

It gives scalable, fault tolerance, reliable, and cost-efficient data storage for big data and it runs on a commodity hardware. Hadoop directly connects with HDFs by shell commands. It is largely a fault tolerant system that is

Revised Manuscript Received on April 18, 2019.

Rahul Shahane, CSE, Koneru Lakshmaiah Education FoundationVaddeswaram, A.P., India
Proddutur Shruthi, CSE, Koneru Lakshmaiah Education FoundationVaddeswaram, A.P., India
J. Sai Ram Prakash,CSE, Koneru Lakshmaiah Education FoundationVaddeswaram, A.P., India.

necessary for data storage on the cluster.

Components of HDFS:

i) Name Node: It is known as master node. It does not store the real data instead it stores the metadata i.e., the number of blocks, their location, rank. It majorly consists of files and directories. The tasks which is done by hdfs are control the file system namespace, supervises the client access to files and performs the file execution like opening, closing files and directories.

ii)Data Node: It is also known as slave and it is in charge of the stocking of the real data in HDFS. This performs read and write operations as per the appeal of the customer. Replica block consists of two file, one is for data and the second one is for recording the block's metadata. If any mismatch is found then the data node goes down automatically. The tasks done by the data node are the operation like block replica creation, deletion under the surveillance of the name node and also manages the data storage of the system.

iii) MapReduce: It provides data pre-processing. It is a software framework for writing applications that process the huge amount of structured and unstructured data which is stored in HDFS. These are very useful for executing huge scale data analysis using so many machines which improve the speed and reliability of the cluster. It has two phases map phase and reduce phase. Map function changes one set of data to another set where the single data is divided into tuples. Reduce function extract the output of the map function as an input and connects the tuples according to their key values. Features of the map-reduce function are simplicity, scalability, speed and fault tolerance. It is a programming technique that is highly potent. It is used for large amounts of data that needs to undergo distributed processing on cluster [1].

iv) Yarn: The full form of YARN is yet another resource negotiator. It gives the resource management and one of the most important components of the Hadoop ecosystem. It is called as the operating system of the Hadoop as it is in the control for managing and monitoring the work loads. It permits various data pre-processing engines to hold the data saved in a single platform. The features of the YARN are flexibility, efficiency and shared.

v) HBase: It is a NoSQL database that is column oriented which is used for Write/Read access. It is designed to store the distributed database which contains billions of rows and columns of structured data. It is scalable, distributed and NoSQL database which is built on the top of the HDFS.

B. Components of HBase

i)HBase Master: It observes and takes good care of Hadoop cluster, performs the operations like addition, deletion, updating the tables. It also controls the failover and manages the DDL operation.

ii)Region Server: It is node which supervises the reads, writes, delete and updates requests from the customer. It runs on the HDFS Data Node.

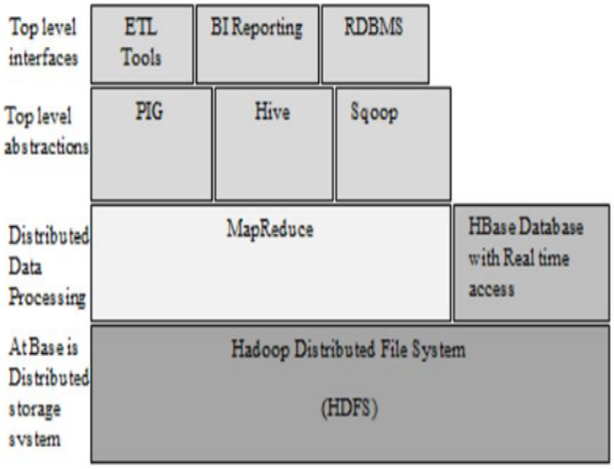
iii) Pig: It is high- level platform for examining and questioning the large data sets which are saved in the HDFS. It uses PigLatin language. It piles up the data and applies whatever filters are required in the format. The features are extensibility, optimization opportunities, handles all sort of data. It enables data workers to write complex data transformations without knowing Java. It examines both structured and unstructured data present in the HDFS.

iv) Hive: Hive provides a SQL-like interface to data stored in HDP. It is an open source data warehouse system for examining and questioning a huge data sets which are stored in Hadoop files. It mainly does three functions- data summarization, query and analysis. It uses a language called HiveQL which automatically converts the SQL-queries to MapReduce jobs which we will get the output in Hadoop. The main parts of the hive are metastore, driver, query compiler, hive server.

v) Sqoop: It will bring the data from the outside sources into Hadoop ecosystem components like hive etc. Similar data could be transferred or imported between database and Hadoop. It works with relational databases such as MYSQL, oracle etc. The features of Sqoop are import sequential datasets from mainframe, import direct to orc file, parallel data transfer, efficient data analysis and fast data copies.

vi) Oozie: A centralized arrangement for Hadoop jobs that could be carried out only by help of work process management. It merges multiple jobs into one logical unit of work. In this the users can create DAG which can run parallel and sequential in Hadoop. It is scalable and is very much flexible. We can easily start, stop, run and suspend the jobs. We can also skip a particular failed node or return it to the oozie.

vii)Zookeeper: This is a centralized service and it maintains the information, naming, provides distributed synchronization. It manages a huge cluster of machines. It is fast and maintains all the records of the transactions in ordered manner.



V. HADOOP SECURITY OVERVIEW

Initially the concept of security was missing in the Hadoop. There are no security hierarchy, no user authenticity verification, information privacy has no scope by any means in conjunction with services so that any provided arbitrary code could produce a result. Hadoop is a framework which is built without a security. It was a tool to run map-reduce problems on large number of data sets. It is a dynamic framework of features, functions, which makes security more difficult. Each has its own module and its own version of the code and some may require independent architecture to work in cluster. Each one

brings its own security options and its issues related to it. [18]

In this environment thedata is processed whenever the resources are handy and is shored up by parallel computation. Hadoop's architecture is highly unsafe to attacksat various points. Data can be sliced into fragments that can be shared across various servers.

VI. THREATS TO SECURITY IN HADDOP

A. Unauthorized client

An unauthorized client means it does not have access to do any operation. By using the pipe line streaming Data-transfer protocol a client is who is not authorised may write/read a file's block of data at a Data node. If a suspicious activity is being cited by performed by an ambiguous user to gain access privileges it can submit a query to a queue, delete, change priority of the query. And it can use query trackers HTTP shuffle protocol to access the intermediator data

B. Task

A task that is being executed can access other tasks or local data containing output of map or locally stored data node using host OS interface.

C. Unauthorized users

An unauthorized user can attack further with accessing the HDFS file via the RPC (remote procedural call) or via HTTP protocols. At the similar time he may sniff/eavesdrop to data packets which are being sent to the clients by data node and can submit a workflow to Oozie as a new user. Data node will not impose access protocol, they can bypass the various access protocol mechanism [10]

VII. SOLUTIONS FOR SECURITY IN HADOOP

A. Authentication

Identifying user or system verification which access the system is called Authentication. Hadoop uses Kerberos as an initial authenticity verification. In early SASL was used in implementing Kerberos and simultaneously authenticates users, their apps, and other Hadoop services over RPC connections [7]. It also supports "Pluggable" verification to HTTP Web Consoles, meaning implementers of web applications and web consoles. HBase with the help of RPC, HTTP provides the much-needed authentication to form a secure client connection which is required for SASL Kerberos. A two-way protocol which is specific on providing authentication called Delegation token is mediated in between authenticating users which are user and name node. It is unsophisticated and is more effective and efficient than that used in Kerberos which is a three-party protocol. [7,15]

B. ACL's & Authorization

The process of laying out access control privileges for system or user is called authorization. At Hadoop, file-based permissions are used to implement access controls, which follow the permissions model of UNIX. in HDFS, Name Node can enforce access control to files by file permission. MapReduce provides ACLs for job [13].

C. Encryption

The data should be secure while transferred into and out of the Hadoop system.In Hadoop ecosystem SASL (simple authentication and security layer) authentication framework is used in the data. SASL security assure of the data exchange taking place between client and servers and guarantee that it cannot be readable by any "man-in-middle". [15]

D. Audit trails

The main purpose of audit is to meet the security compliance requirements,log monitoring has become an essential ingredient in the system in order to gain frequent interval basis on the entire Hadoop ecosystem. Base audit support is used by HDFS and MapReduce. The security of the information is the most important for the companies to have victory journey in securing the big data. There is always a possibility to have violation in the security which will be caused by the unauthorized and unintended access by some users. So, to meet this problem we need to audit the entire Hadoop system and deploy a system of log monitoring.

VIII. TECHNOLOGIES USED FOR SECURING HADOOP

A. Zetta set orchestrator

To use the different fragments in Hadoop, a great product named Zetta set Orchestrator is used [6]. It does consist of highly compiled set of metadata as a repository which serves as integral part in GUI together with workflow. The data is said to be prevented and served in close relation with data, which is served equally on basis with prime-time security devices could fail at times. [6]

B. Apache sentry

It is a part of open source software from cloud era. It is a module used for authorising Hadoop which offers the granular. It works for both role-based authorization and fine-grained authorization. It is required to provide accurate level of access to the correct user and applications. [11,15].

C. Apache Knox

It is the gateway of the system which provides single point of authentication and accessibility to multiple servers in the Hadoop cluster. It provides the security to different versions across various servers of Hadoop.The authentication type is single point of verification and for accessing various types of Hadoop services in cluster [9]. It provides outline security solution to Hadoop.

Advantages: It provides support to various verification and token authentication scenarios. [9]

D. Project Rhino

Project Rhino gives an integrated point to point data security to the Hadoop ecosystem. Token based authentication and SSO solution are provided by Project rhino. It supports the key distribution so that MR can decrypt the data block and give the result as per the requirement needed. By this the security of the HBase is increased because of the cell level authentication and encryption for table stored in Hadoop. It uses a block level encryption.

	Authentication type	Authorisation type
Apache Sentry	-	Granular
Apache Knox	Single point	-
Project rhino	Token Based	-

IX. CONCLUSION

Today data usage and size are increasing rapidly, in this era of big data where the origin of data cannot be clearly demarcated, and we need to secure data coming from various sources. To process this data Hadoop is used, in this paper we have made a study on various security issues in Hadoop environment and the possible solution is implemented.

REFERENCES

1. Cloud Security Alliance "Top Ten Big Data Security and Privacy Challenges"
2. Tom White O'Reilly |Yahoo! Press "Hadoop the definitive guide"
3. Owen O'Malley, Kan Zhang, Sanjay Radia, Ram Marti, and Christopher Harrell "Hadoop Security Design"B. Smith,
4. Mike Ferguson "Enterprise Information Protection - The Impact of Big Data
5. Vormetric "Securing Big Data: Security Recommendations for Hadoop and NoSQL Environments, October 12, 2012".
6. Zetta set "The Big Data Security Gap: Protecting the Hadoop Cluster"
7. Devaraj Das, Owen O'Malley, Sanjay Radia, and Kan Zhang "Adding Security to Apache Hadoop"
8. Seref SAGIROGLU and Duygu SINANC "Big Data: A Review Collaboration Technologies and Systems (CTS), 2013 International Conference, May 2013"
9. Horton works "Technical Preview for Apache Knox Gateway"
10. Kevin T. Smith "Big Data Security: The Evolution of Hadoop's Security Model
11. M. Tim Jones "Hadoop Security and Sentry"
12. Victor L. Voydock and Stephen T. Kent "Security mechanisms in high-level network protocols. ACM Comput. Surv.1983".
13. Vinay Shukla s "Hadoop Security: Today and Tomorrow"
14. MahadevSatyanarayanan "Integrating security in a large distributed system.ACM Trans. Comput. Syst., 1989"
15. Sudheesh Narayana, Packt Publishing "Securing HadoopImplement robust end-to-end security for your Hadoop ecosystem"
16. S. Singh and N. Singh, "Big Data Analytics", 2012 International Conference on Communication, Information & Computing Technology Mumbai India, IEEE, October 2011
17. jeffhurtblog.com "three-vs-of-big-data-as-applied-conferences, July7,2012" Priya
18. Big Data Storage in Hadoop: A Review of Security Issues and Threats Kavita K. Kanyan, Er. Ritika Mehra

AUTHORS PROFILE

Rahul Shahane, Asst Prof . KLEF Deemed to be University, Guntur , AP. As well as Phd research Scholar at KLEF. Major area of research is Deep

Author-2 Photo

Proddutur Shruthi, personal persuing Btech CSE from KLEF Deemed to be University. Area of interest is Big Data and Hadoop.