Volume 8, No. 5, May-June 2017
ISSN No. 0976-5697
International Journal of Advanced Research in Computer Science
REVIEW PAPER
Available Online at www.ijarcs.info

Design of Mechanism for Enhancement the Security of Hadoop Processing Tool "Hive" With VMWARE Platform

Anjali Devi
Master of Technology, Department of Computer Science & Engineering,
Bhagat Phool Singh Mahila Vishvavidyalya,
Khanpur Kalan, Sonipat, Haryana, India

Abstract : Hadoop is a completely open-source software background for loading the data and make running environment for the applications on commodity cluster with the computing. It offers huge storage for various data, huge processing authority and the capability to hold virtually unlimited parallel tasks or jobs. The logs are structured, unstructured, semi-structured model generating in gigantic volume of information and cannot handle using RDBMS easily. Accordingly, Hadoop master-slave cluster is used, this cluster controlled upon the largeinformation. Consequently, whole process includes data import from servers to HDFS then all the processing done step by step in the frequent manner. The underlying framework HDFS architecture with map reduce shall be definable using HIVE with Kerberos authentication algorithm, generating reports using core map reduce and processing through VMware Cloudera management with Hadoop. the information extracted and convert to more interactive and understandable bar graph which can be improve their meaningful data and existing units.

Keywords*:* Bigdata, Hadoop, HDFS, MapReduce, HIVE, security, VMware, Kerberos.

1.INTRODUCTION

Nowaday, we alert of the stage of information security, where all that backgrounds link to a bulky information than constantly before. Now the data measured with The Volume – challenging to "load and process" in this increase the size of data day by day.Variety – some "data types thenstoringbuilding" of the suggestion. Velocity –actual processing influenced by rate of data entrance in velocity work will be done through "Batch to streaming data". Veracity- confirming inference-based models aftercomplete data groups. Variability- responsibilities to absence of consistency and safe the policy. Venue – location terminology. Now, data security play very vital role so problems also increase continually, however new utensils and technologies are growing. So, the Hadoop using HDFS as a storage and MapReduce as a processing with ecosystem tool like HIVE with Kerberos authentication procedure is an excessive stage to creation Hadoop atmospheres protected [1].

2.HADOOP SECURITYTECHNOLOGY TIMELINE

Hadoop timeline: - 1999-APACHE software foundation work formed as a non-profit to make the Hadoop as an open source. In 2002-Nutch created by Doug Cutting and Mike Cafarella. The HDFS is formed on founded on 2003 google present Google File System(GFS), white paper.
2004 processing framework map reduces basic simplified data processing on huge clusters. In 2006Hadoop 1.x version is released Doung Cutting Joins Yahoo, take Nutch with him. 2008 – Nutch divided and Hadoop is born with the VMware +cloudera platform. 2010- HIVE and PIG graduates. 2013- YARN deployed atYahoo. Common includes Java reference library and services essential by further Hadoop modes. These libraries provide filesystem and OS level notions and contains the compulsory Java files and scripts required to start. Hadoop distributed File System 2. x yarn framework for the task arrangement and cluster source administration.

Firstly, there was no security prototypical – Hadoop didn't authenticate employers or services, and there was no data secrecy. As Hadoop was calculated to execute cypher over a distributed cluster of machines, anyone might submit code and it would be implemented. Although auditing and authorization panels (HDFS file approvals) were implemented in earlier distributions, such access control was easily avoided because any user could copy any other user with a facility line switch. Though, since there stood insufficient security panels within Hadoop, various chances and security occurrences chanced in such surroundings. Hadoop subproject created with mailing lists and Wikipedia [2].

3. STORAGE INFRASTRUCTURE AND PROGRAMMING MODEL: MAPREDUCE

It is moderately classy to build superior servers with heavy arrangements that handle bulky scale processing. The Hadoop consuming two chief mechanisms Hadoop Distributed File System then the MapReduce family. HDFS is used to split the data and store on a separate node within the cluster, Information is originally separated into manuals and records. Files are divided into continuous sized chunks of 128M and 64M preferably128M MapReduce is used to process the information.

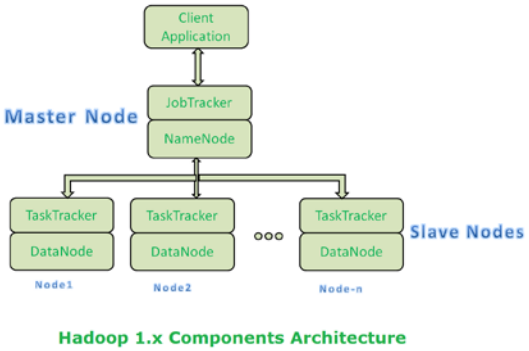

Figure No 1: Basic Block Diagram of Hadoop Core Components [3]

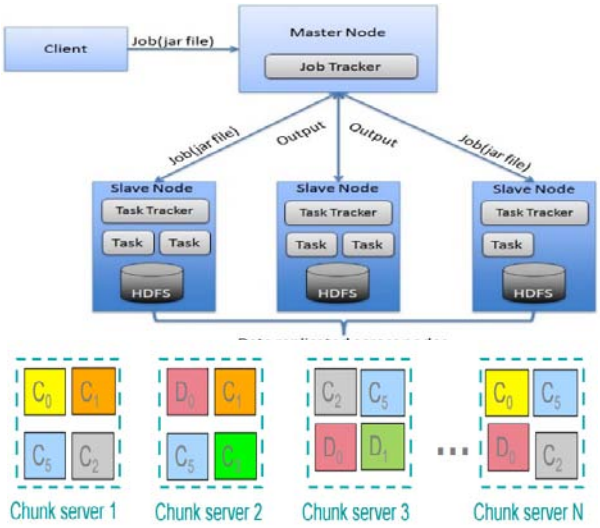

Figure No 2: Programming Model ofHadoop processing framework [4]

A. HDFS

- An HDFS collection has two categories of nodes functioning in a master-employees pattern: AName Node machine (the master) and a capacity of data nodes (employees).
- The Name Node machine achieves the file scheme. It saves the filesystem order and the meta message for all the records and directories in the hierarchy.
- The namenodemachine also recognizes the data nodes on which all the chunks of an assumed file are positioned; though, it prepares not store block locations determinedly, as this data is recreated from the data nodes after the scheme begins.
- Data nodes are the supports of the file scheme. They stock and recover chunks when they are expressed to, their description backbone to the Name Node machine sometimes with lists of chunks that they are storing.
- Deprived of the Name Node machine, the file arrangement cannot be used. In fact, if the mechanism successively the Name Node machine was smashed, all the files on the case system would be lost since there would be no way of expressing how to reconstruct the files from the chunks on the data nodes
- For this motive, it is dynamic to change the Name Node mechanism, machines healthy to dissatisfaction, and Hadoop delivers two mechanisms for this.
- The first machine is to backbone the cases that face the determined state of the case system metadata. Hadoop bottle be planned so the Name Node mechanism writes its strong-minded state-run to some case systems. These writes are coordinated and atomizers. The standard plan excellent is to write to native disk as fine as an isolated NFS gateway base [5].

B. MapReduce v1

MR is a simple software programming model built on Java language, used for script submissions to process huge quantities of information, in equivalent, on huge clusters of low-priced hardware, it is consistentsystem work without losing the information, it abstracts the trouble of distributed allowance and parallelism from developer, Computation is taken with nodes where information exists on the machines. This is labelled 'dataLocality'. Map reduce algorithm breaking the allowance into three phases:

MapperEach Map task usually operates on single block input split in HDFS.

Shuffling and Sorting Segment This phaseshuffle and sort the information according to the mapper output.

ReducerOperates on shuffle/ sort intermediate data i.e. Map responsibilities of shuffle /sort output.Combines all the results of mapper shuffle /sort and reducer and Produces final output.

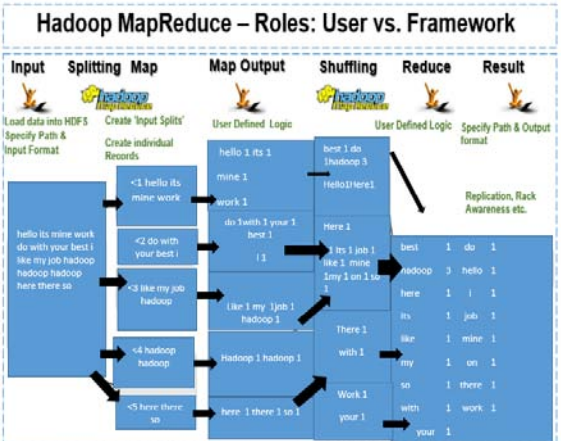

Figure No 3: Map Reduce Algorithmic Output

C. MapReduceV2 (YARN)

Yet Another Resource Negotiator is used to manage the cluster management. it is an advance version of Hadoopi.e. 2.x,described as a redesigned resource manager[6].

D. HIVE

Data Warehouse set-updocumented by Facebook.It's offers SQL-like language: HiveQueryLanguagesupplementary combination with Hadoop security [7].

E. Use Case

An encryption, sector is an encyclopaedia in HDFS with completely its contents, that is, separate file, subdirectory, encoded. The records in this directory determination be clearly encrypted upon to write and clearly decrypted upon read. Each encryption neighbourhood is connected by a key which is fixed when the district is formed [8].

F. Key Management

Respectively file within an encryption zone also has its individual encryption/decryption key, named the Data Encryption Key (DEK). Anoriginal deal requests to be additional to your cluster to stick, succeed, and access encryption keys, called the Hadoop Key Management Server (KMS). The KMS facility is a representation that boundaries with a support key store happening behalf of HDFS daemons then clients. Together the support key store and the KMS tool the Hadoop Key Provide customer API. [9]

G.Secure Sockets Layer

Cloudera strongly mentions in contradiction of the use of self-signed documentations. The key store must cover an effective certificate. The Hadoop community understood that additional strong security panels where compulsory emphasis on authentication, and selected Kerberos as the authentication machinery for Hadoopsecurity. [10]

4.TODAY'S APACHE HADOOP ENCRYPTION SECURITY ARCHITECTURE

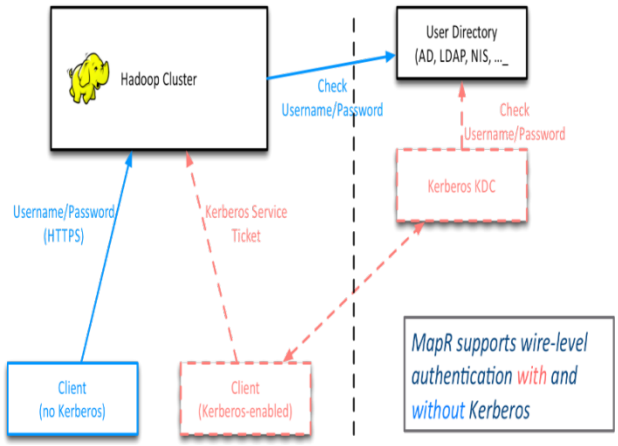

Figure No 4: Security Architecture [11]

Step 1: -Basic Data Encryption Standard

- In the encryption process transposed and split the input 64 -bit into Right half and Left half.
- Gets a left shift and is transposed (56- bit key)
- Substitution method is used to modify the right half key
- One more round of transposition takes place
- Left half and the modified right half combine to form the new right half
- Old right half becomes the new left half
- This cyclecontinues 15 more times

Step 2: -The Hadoop Encryption

- Firstly, data key generated, generate the key.
- Map Reduce job runs on the cluster and generates data.
- Data need to be encrypted with clusterconfiguration.
- Data key used to encrypt data as output.
- Encrypted data and data key stored.
- Key- encrypting key stored separately.

Step 3: -Data Node Uses Symmetric Key to Decrypt the Data Block

- Client requests encrypted data to name node. Trust stores with certificates and key store with symmetric keys client -side certificates and symmetric key are used for data access.
- Keystone with symmetric keys and public key pairs /certificates.
- Namenode authenticates the request using its own key stores to compare symmetric Key.
- If the mostsignificantis abinding name node carries a list of nodes property the information.
- Key stores with symmetric keys and public keypairs/ certificates,client requests data blocks of encrypted data to data node.
- Data node uses the key to decrypt the data block and if successful passes it back. also, respective data nodes pass subsequent data blocks back.
- Data node communicates to decrypt a retrieve subsequent data block.

Step 4: - Data Encryption at IntelHadoop Distribution*Using HIVE*

HDFS encryption by consuming all Hive info inside the similar encryption , region. Cloudera Manager , The Hive Scratch Directory to be private encryption region.

- Create symmetric key and keystores.
- Create a key pair (private /public) and key store.
- Create a trust store contain public Certificates.
- Extract certificates from a key store define in step 2 and import them into a trust store.
- Hadoop component "pig" uses the symmetric key to encrypt hdfs file.
- Hadoop component "HIVE" defines an encrypted external table uses the symmetric key created in step 1.
- Authorized clients access encrypted data through map reduce jobs using certificates from trust store[12].

5. KERBEROS

Kerberos is a system authentication procedure. It is strategic to distributestrong authentication for client/server requirements by using secret-key cryptosystem. This authentication machinery is presented only for Hive Server distributions.

The algorithm of Kerberos authentication is as following: -

Phase 1: Kerberos authentication is founded on symmetric key cryptography to validate the information. In symmetric significant cryptography, the interactive objects use the similar key for mutually encryption and decryption with keys.

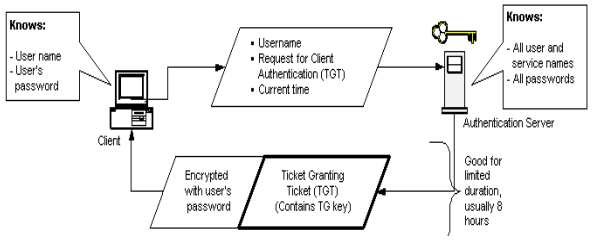

Figure No 5.1: Initial step-upfor aticket granting and authentication is based on symmetric key

Phase 2: The Kerberos key distribution center (KDC) delivers scalability. The Kerberos procedure continuously contracts with encrypted ticket granting key, service request, time limitation.

Phase 3: A Kerberos ticket deliverssafeconveyance of a session key.

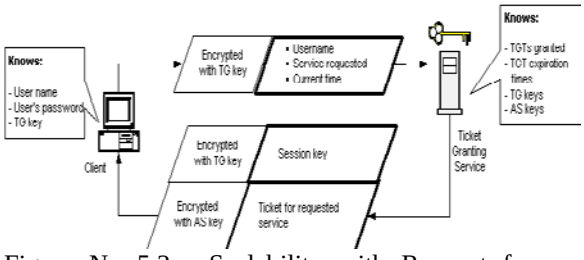

Figure No 5.2 : Scalability with Request for servicing, Session Key Management

Phase 4: The Kerberos KDC allocates the meeting key by distribution it to the customer.

Phase 5: The Kerberos License Granting Ticket boundaries the usage of the entities'principal keys.

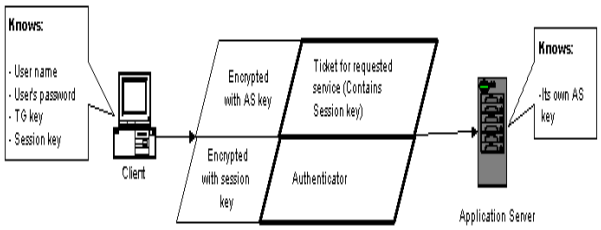

Figure No 5.3: Communication between client and the application server

Phase 6: Application Server Reply to Alice with another resource authenticator: Alice sends the TGT request , then server , send the request accepted message with TGT and session key , then Alice reply with request ticket + authentication note to the server the ticket second-hand to validate to the supply server, and binary session keys, after that the Alice request to the resource server with authentication , one to authenticate to the KDC and another to authenticate to the resource server [13].

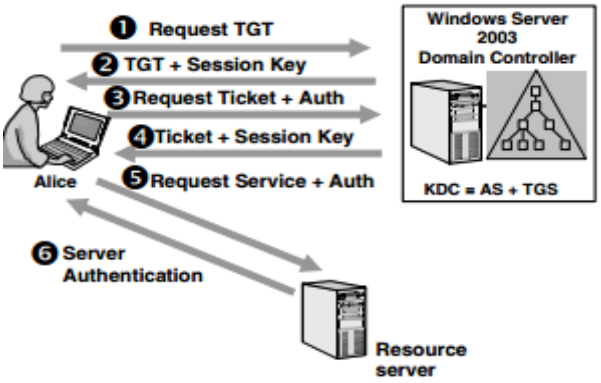

Figure No 5.4: The complete Kerberos protocol

6. VMWARE WORKING AREA

The role of set-up, whether it's physically or else, virtually, is to support requests. This includes traditional professional critical submissions as well as modern cloud, mobile as well as big data submissions Virtualizing biggest data applications like Hadoop security proposals a lot of assistances that cannot be found on physical structure or in the cloud. Shortening the management of your big data assembly becomes quicker time to consequence, creation it extra cost effective. VMware is the top stage for big data just as it is for old-style requests.

VMware's Role in Big Data

- **Simple**
  Simplify procedures and care of your big data structure.
- **Agile**
  Get your structure on request so you container rapidly distribute business world.
- **Flexible**
  Multi-tenancy permits you to track various Hadoop distributions on the similar virtual machine.
- **Efficient**
  Modernization-of capacity, mobility adds to processing efficiencies.
- **Secure**
  Ensure control and submission of your complex data. [14]

7. CONCLUSION

In the information era, we are currently living in, voluminous varieties of high velocity data are presence formed daily, and within them in expert vital and plans of hidden information which should be extracted and utilized. Hence, big data analytics can be applied to leverage business change and enhance decision manufacture, through applying progressive analytic methods aimed at big data, and tight fitting hidden visions and valuable knowledge. Accordingly, an analysis of the big data analytics concepts which are being researched, as well as their importance to decision making with the applications like VMWARE machine. Consequently, big data were discussed, HDFS, Map reduce and HIVE are important. We can split the Kerberos protocol into three main ladders: Authentication process, where the user (and host) get a Ticket Granting Ticket (TGT) as authentication token, Service request process, where the user obtains a Ticket Granting Service (TGS) to access a facility, Service access, where the user (and host) use TGS to authenticate and access a detailed service. Moreover, some of the biggest information, analytics devices and the approaches were examined. IN addition, some of the different advanced data analytics techniques and security of information were further discussed.

REFERENCES

[1] https://www.slideshare.net/markogrobelnik/big-datatutorial-grobelnikfortunamladenicsydneyiswc2013
[2] https://en.wikipedia.org/wiki/Apache_Hadoop
[3] http://www.journaldev.com/8808/hadoop1-architecture-and-how-major-components-works
[4] http://www.mmds.org/mmds/v2.1/ch02-mapreduce.pdf
[5] https://www.youtube.com/watch?v=qV__qpP8NDo
[6] http://www.seminarstopics.com/seminar/653/hadoop-technologies/
[7] http://infolab.stanford.edu/~ragho/hive-icde2010.pdf
[8] https://www.cloudera.com/documentation/cdh/5-0-x/CDH5-Security-Guide/cdh5sg_encryption_enable.html#topic_14_unique_2
[9] https://www.ibm.com/support/knowledgecenter/SSPT3X_3.0.0/com.ibm.swg.im.infosphere.biginsights.admin.doc/doc/kerberos_hive.html
[10] https://www.slideshare.net/cloudera/protecting-hadoop-data-at-rest-with-hdfs-encryption
[11] http://www.idevnews.com/stories/6008/MapR-Boosts-Hadoops-Out-of-the-Box-Security-with-Native-Authentication-Authorization
[12] https://www.cloudera.com
[13] http://search.iiit.ac.in/cloud/presentations/28.pdf
[14] http://www.vmware.com/in/solutions/big-data.html