Challenges and Security Issues in Implementation of Hadoop Technology in Current Digital Era

Dr. Vinay Kumar, Ms. Arpana Chaturvedi

Abstract—With the advent of technologies, managing tremendous amount of over flown and exponentially growing data is a major area of concern today. This is particularly in terms of storing and organizing data with security. The exponentially growing data due to Internet of Things (IoT) has led to many challenges for the governmental and non governmental organizations (NGOs). Security threats forced to the private and public organizations to develop their own Hadoop based cloud storage architecture .In Apache Hadoop architecture it creates various clusters of machines and efficiently coordinates the work among them. Hadoop Distributed File System-HDFS and Map Reduce are two important components of Hadoop. HDFS is the primary storage system used by different applications of Hadoop.It enables reliable and extremely rapid computations. HDFS provides rich and high availability of data to different user applications running at the client end. Map Reduce is a software framework for analyzing and transforming a very large data set into desired output. This paper focuses on the review of HDFS 0, HDFS 2.0 and HDFS 2.8 architecture, and its various functionalities including analytical and security features.

Index Terms—Cloud Computing, Clusters, Hadoop, HDFS, Hive, IoT, Map Reduce Pig, Sqoop.

1 INTRODUCTION

Hadoop is an open source architecture which is used to store the structured, semi structured, unstructured, quasi structured data ,collectively such data is termed as big data.It provides meaningful output using data analytics. The standard process used to work with big data is ETL (Extract, Transform and Load).Extraction means getting data from multiple sources, Transform means convert it to fit into analytical needs and Load means getting it into the right systems to derive meaningful value out of it. It provides various benefits to governmental as well as non governmental organizations. The collected data is of two types, operational data and analytical data. The different types of data comes under two categories are: **Transactional data**, generated from all daily transactions**,** **Social Data**-generated from different social networking sites like Face book, Google ads etc. **Sensor or Machine Data**- generated by industrial equipment, sensors that are installed in machines, data stored in black box in aviation industry, web logs which tracks the user behaviors, medical devies, smart meters, road cameras, satellite, games and many more Internet of Things .All Government organizations are now-a-days getting digitized and aadhar enabled.Aadhar enabled applications will provides better services and facilities to the right person as an individual and let the citizens participate in digital economy. To implement digitization in different organization and to utilize all the benefits now-a-days companies are moving towards Hadoop technology from existing one.Hadoop is a highly scalable platform developed in JAVA, which consists of distributed File system that allows multiple concurrent jobs to run on multiple servers splitting and transferring data and files between different nodes. It is efficient to process or recover the stored data without any delay in case of failure of any node. At the same time chances of fraudulence increases while processing or storing information in HDFS.Due to various big data issues with respect to management, storage, processing and security, it is necessary to deal with all individually [8].

This paper is organized into five sections.Secion 2 deals with literature review. Hadoop File system, its architecture and components are discussed in section 3. Existing problem and the challenges are outlined in Section 4 and paper is finally concluded with the proposed solution in the section 5.

- *Vinay Kumar is a Professor in Vivekananda Institute of Professional Studies, Delhi. Earlier he worked as Scientist in NIC, MoCIT Government of India. He completed his Ph.D. in Computer Science from University of Delhi and MCA from Jawaharlal Nehru University, Delhi.He is member of CSI and ACM. Ph: 011-2734 3402. E-Mail:vinay5861@gmail.com*
- ***Arpana Chaturvedi** is working as an Assistant Professor in Jagannath International Management School, Delhi. She is M.Sc. (Math), MCA and M. Phil. (Comp. Sc). She is pursuing PhD from Jagannath University. PH-01149219191. E-mail: ac240871@gmail.com*

2 Review and Existing Problem

2.1 Lierature Review

J.Zhao, L.Wang, J Tao, J. Chen, W. Sun, R. Ranjan has suggested that Map Reduce is viewed as the appropriate programming model for extensive scaled information based applications [1]. Hadoop based system uses map reduce programming to run on different clusters.G-hadoop reuses the client validation and occupation accommodation system of Hadoop, which is intended for the solitary group. They proposed security model for Hadoop which depends upon open cryptography and SSL convention. This security structure opens up the client's confirmation and employment accommodation procedure of the present G-Hadoop execution with a solitary sign-on methodology [2].V. Kadre, Sushil Chaturvedi proposed AES-MR encryption scheme for securing Data in HDFS Environment .The AES encryption algorithm is one of the best approaches to encode data. It works in parallel. They broadly utilized IEEE 1619-2007 standard XES-TCB-CTS(XTS) mode in which key material for XTS-AES comprises of encryption key. The XTS mode permits parallelization and pipe lining which empowers the last deficient piece of data.[3]Monika Kumari, Dr. Sanjay Tyagi suggested three layered security model for data management in Hadoop environment. In this approach a secure tunnel based transmission is provided for communication with authenticated users. One time authentication is provided by RSA algorithm, SSL layer is activated to avail Hadoop services. For free users RSA based authentication is performed to allow public area access. The security is implemented in the middle layer which is divided into 3 parts, Authentication, Secure Session and Secure Data management. [3]Rajesh Laxman Gaikwad,Prof. Dhananjay M Dakhane ,Prof Ravindra L Paridhi has proposed Network security enhancement in Hadoop Clusters by introducing automation in authentication using Delegation tokens and suggested advanced security models in the form of Security Enhancement and security using Role Based Access Control with discussion about developments in Web authentications for Internet-based users of Hadoop Clusters.

2.2 Problem to be discussed in this paper

All existing and growing private and government organization are adopting Hadoop based cloud storage architecture. All crucial and personal data will be lying in the storage architecture of Hadoop. It keeps sensitive information in multiple nodes, clusters or servers in the form of separate files. It uses so many technologies Hive, Pig, HBase, and Mahout to analyze data more efficiently and effectively. Most of the private and government organizations have fear in keeping their data in Hadoop [13].Hadoop has no Security feature implemented by default, which later on arise so many security related issues like misuse of personal data and fraudulent issues. At the time of Hadoop implementation one should ensure that the security features should be implemented in such effective way that only authenticate user should be able to use data, no case of fraudulent or misuse of information should arise.

Challenges of Hadoop:

It has many challenges which are to be overcome so that all organization can rely on it and store ever-growing data into it with reliability and security. At present it has following challenges:

1) Constant growth in data: As the data is ever growing and exponential, the Hadoop clusters are also need to be scaled. Its ecosystem consists of complex set of software, which keeps on changing as per the demand and necessity in maintaining datasets. The existing scenario has lack of protocols or guidance which can provide the best platform to run it safely.
2) No fix Platform to work on: The Hadoop Community is not having a fixed platform; it depends upon the end user to choose as per the requirement. At the same time end user may not have appropriate knowledge of hardware to provide the best possible solution of the problem.

In 2010 The Economist asserted that data has become a factor of production, almost on par with labor and capital.IDC predicts that the digital universe will be 44 times bigger in 2020 than it was in 2009, totaling a staggering 35 zettabytes. EMC reports that the number of customers storing a petabyte or more of data will grow from 1,000 (reached in 2010) to 100,000 before the end of the decade. By 2012 it expects that some customers will be storing Exabyte's (1,000 petabytes) of information. In 2010 Gartner reported that enterprise data growth will be 650 percent over the next five years, and that 80 percent of that will be unstructured.

Hadoop Components and Architecture:

Apache Hadoop software library can detects and handles failures at the application layer hence deliver high availability services in the top of all multiple clusters of computers and make individually each of them less error prone.
Hadoop architecture consists of not only Hadoop components but also an amalgamation of different technologies that provides immense capabilities in solving complex business problems, government projects.

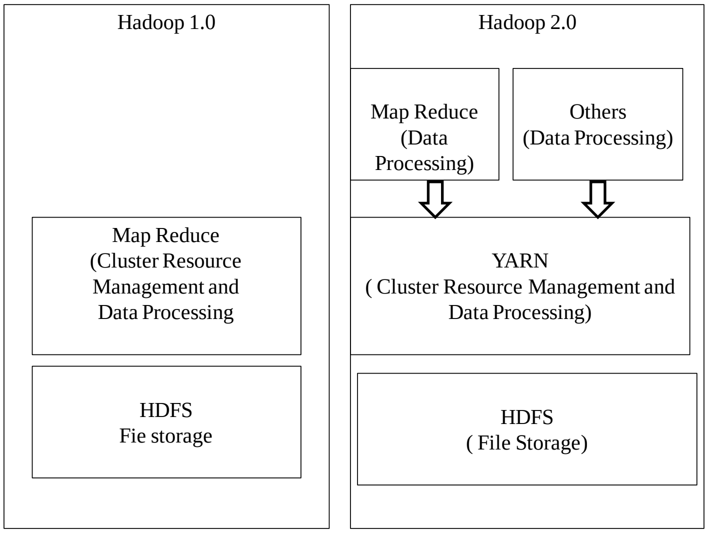

Fig. 1 Difference in Hadoop 1.0 and Hadoop 2.0

On the basis of working of all the components of the Hadoop ecosystem; it has been divided onto five levels. These are:

- Core Components
- Data Access Component
- Data integration Component
- Data Storage Component
- Monitoring, Management and Orchestration Components

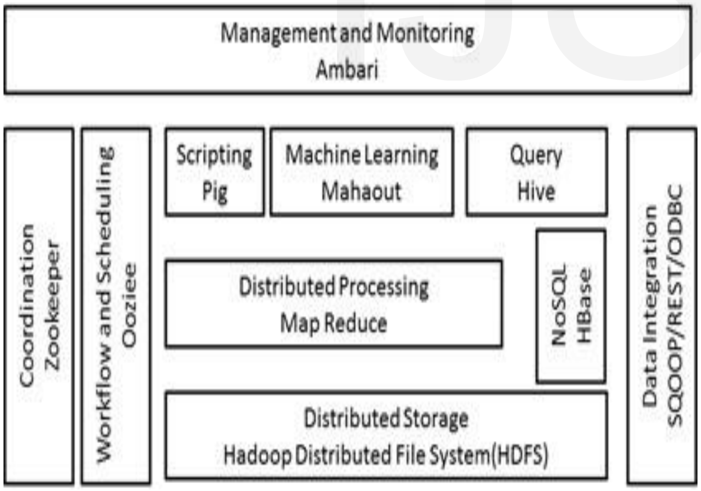

Fig 2. Components of Hadoop

Core Hadoop Components

The Core Components of Apache Hadoop Ecosystem which forms the basic distributed Hadoop framework are comprises of 4 components Hadoop Common, HDFS, Map Reduce and YARN [4].

1) Hadoop Common-

It consists of pre-defined set of utilities, libraries that are used by other all modules exists within the Hadoop Ecosystem. For E.g. HBase and Hive need to make Java archive (JAR) files i.e. jar files, stored in Hadoop common to communicate with or access HDFS.

2) Hadoop Distributed File System (HDFS) –The HDFS, default storage layer is based on Master-Slave architecture model where the Name Node acts as the master node and Data Node acts as a Slave Node. The Master Node i.e. Name Node keeps the track of the storage cluster and the Slave Node i.e. Data Node is responsible to sum up the various systems within a Hadoop cluster.

3) Map Reduce- Distributed Data Processing Framework of Apache Hadoop

Java based Hadoop's Map Reduce is parallel processing system based on Yet Another Resource Negotiater (YARN) architecture. Map Reduce takes care of scheduling jobs, monitoring jobs and re-executes the failed task. The delegation tasks of the Map Reduce component are tackled by - Job Tracker and Task Tracker as shown in the image below –

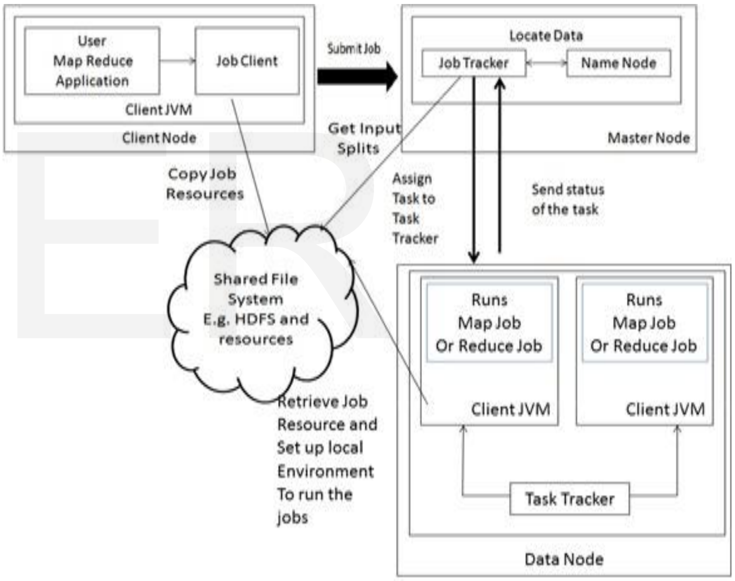

Fig 3. Delegation task of Map Reduce Component

4) YARN

Yet Another Resource Negotialter (YARN) introduced in Hadoop 2.0 is a dynamic resource allocator in the Hadoop framework as users can run various Hadoop applications without having to bother about increasing workloads.

2) **Integration Components with Databases or Data Access Components used by Enterprises:** The other data access components of Hadoop Ecosystem forms an integral part of Hadoop Ecosystem, enhances the strength of it as provide better integration with databases, makes Hadoop faster with new features and functionalities. These eminent Hadoop components are Pig and Hive. **Pig-** Apache Pig provides optimized,

extensible and easy to use high level data flow language Pig Latin. It is developed by Yahoo for analysing voluminous datasets efficiently and easily. **Hive**-It uses HiveQL language which is similar to SQL for querying and analysing the data. It was developed by Face book. It can summarize data from data warehouse and makes query faster through indexing.

Data Integration Components of Hadoop Ecosystem- Sqoop and Flume

Sqoop: It is used for import and export purpose both. It imports the data from external sources into related It also exports data from Hadoop to other external structured data stores. It copies data quickly, performs efficient and faster data analysis as can transfer data in parallel and also mitigates excessive loads.

Flume-It is used for collecting data from the source as gathers and aggregate voluminous data and stores it back to HDFS.It can perform it properly by outlining data flows which consists of 3 primary structures channels, sources and sinks. The processes that run the dataflow with flume are known as agents and the bits of data that flow via flume are known as events.

Data Storage Component of Hadoop Ecosystem –HBase
HBase –

HBase is a column-oriented database that uses HDFS for underlying storage of data and helps NoSQL database enterprises to create large database with millions of rows and columns. It is best to use when random read and write access are required to access large datasets as it supports random reads and batch computations using Map Reduce.

Monitoring, Management and Orchestration Components of Hadoop Ecosystem- Oozie and Zookeeper

- **Oozie-**It is a workflow scheduler that runs on java servelets container Tomcat where the workflows are expressed as Directed Acyclic Graphs. It manages all Hadoop Jobs like Mapreduce,Sqoop,Hive and pig as stores all running workflow instances, their states and variables in the database which are executes on the basis of data and time dependencies.
- **Zookeeper-**

Zookeeper works as coordinator as responsible for synchronization service, distributed configuration service and for providing a naming registry for distributed systems hence provides simple,fast,reliable and ordered operational service for a Hadoop cluster.

The other components of Hadoop Ecosystem –

The other common components of Hadoop Ecosystem are: Avro, Cassandra, Chukwa, Mahout, HCatalog, Ambari and Hama. The user can provide appropriate solution to the requirements of any business organization or to government sector to perform big data analytics by implementing one or more Hadoop components.

Need of Hadoop:

With the advent of technology and implementation of it in the form of digitization exploded the huge amount of structured and unstructured data with the increasing volume day by day. It has increased the demand of high storage capacity, management of information, accessing it, analyzing it and the need of management of this data with security so that it can be analyzed and extracted without any loss of information. All the organizations are moving the data on Hadoop architecture because of the following special features it has:

1) Capability of storing and Processing Variety of Complex Datasets in distributed Systems.
2) Fast and Reliable parallel and multiple node Computational ability at the CPU cores.
3) Fault Tolerance and High Availability, Ability to handle real time node failures and redirecting to other nodes to handle it at the application layer.
4) Storing and retrieving enormous data at once without data pre-process.
5) Scalable in nature as able to increase in size from single machine to thousands of servers.
6) Servers can be added or removed from the clusters dynamically without any interruption in operation.
7) Cost effective as Hadoop is an open source technology.
8) Compatible in all platforms as based on Java.

The Benefits of HDFS There is little debate that HDFS provides a number of benefits for those who choose to use it. Below are some of the most commonly

- Built-In Redundancy and Failover HDFS supplies out-of-the-box redundancy and failover capabilities that require little to no manual intervention (depending on the use case).
- The hardware and infrastructure if not properly managed can run into the millions. This is where HDFS comes as a blessing since it can successfully run on cheap commodity hardware.
- The characteristics that Big data comprise of data velocity, veracity, value, variety, and volume and its providing access to streaming data[11].
- Portability any tenured data professional can relay horror stories of having to transfer, migrate, and convert huge data volumes between disparate storage/software vendors.
- Scalability is the biggest strength of HDFS as can store data in much more than zeta bytes and retrieves easily on demand
- Moving computation rather than data and providing extreme throughput

The Benefits of Map Reduce: Map Reduce is the data processing engine of Hadoop clusters deployed for Big Data applications. The basic framework of a Map Reduce program consists of the two functions the Mapper and Reducer. These two pillars of Map Reduce can ensure that any developer can create programs to process data that is stored in a distributed file environment like HDFS [5].

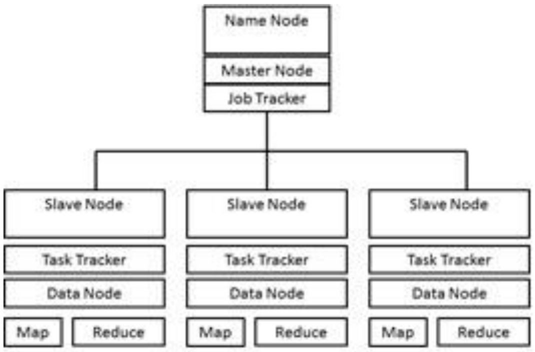

Fig 5. Process of Data using Map Reduce

There are some distinct advantages of Map Reduce and we-have listed some of the most important below:

- Highly economical
- Flexible for multitudinous data
- Extremely fast processing
- Extreme Scalability
- Heightened resilience
- Highly secure system
- Programming simplicity

Proposed Model to Implement Security: In the Security layer I propose to implement security features mentioned in this paper using various techniques. The proposed model is:

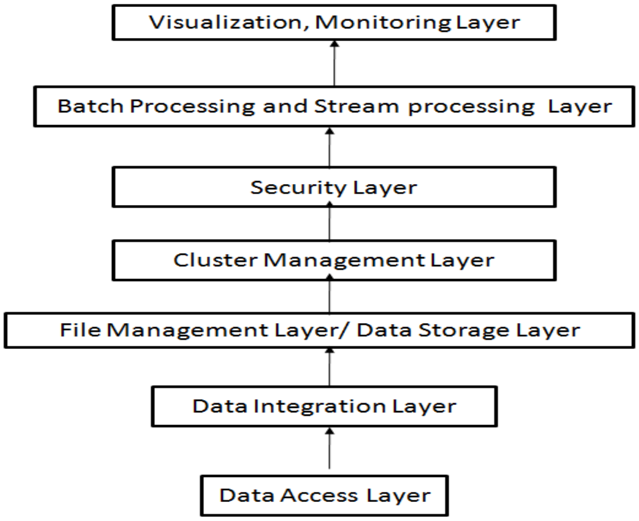
Visualization, Monitoring Layer
Batch Processing and Stream processing Layer
Security Layer
Cluster Management Layer
File Management Layer/ Data Storage Layer
Data Integration Layer
Data Access Layer

Fig 6. ProposedModel for security implementation in Big Data

Proposed Features to handle Big data Security Challenges:

1) **Sharing and Privacy:** There are several different integration models. The idea for big data security analytics is to store more critical or sensitive data in cluster within a cluster using various available data mining technique.
2) **Data Encryption:** This is an important feature to make the big data more secured to access only with the administrator access rights. It has recommended File/OS level encryption because it scales as you add nodes and is transparent to NOSQL operations.
3) **Authentication and Authorization:** To ensure that secure administrative passwords are in place and those application users must authenticate before gaining access to the cluster. Each user has different types of accessing password e.g. developers, users and administrator roles should be segregated.
4) **Node Authentication:** There is a little protection from adding unwanted nodes and applications to a big data cluster, especially in cloud and virtual environment where it is a trivial to copy a machine image and start a new instance. Tools like Kerberos help to ensure rouge nodes don't issue queries or receive copies of data [10].
5) **Key Management**: Data encryption is most important as a key security. Any eternal key management systems are to have secure keys and if possible help validate key usage.
6) **Logging:** Logging is built into Hadoop and any other clusters. It seems to provide the security to all other network devices and applications and recommend that user built-in logging, or leverage one of many open-source or commercial logging tools to capture a subset of system events.

7) **Network Protocol Security:** Secure Sockets Layer (SSL) or Transport Layer Security (TLS) is built-in or available on most NoSQL distributions. It is required to implement protocol security to maintain privacy of informaion and to keep data private.

Implementation in Security Layer:
The Advance Encryption Algorithm (AES) is better thanData Encryption Standard (DES) and Ron Rivest, Adi Shamir and Leonard Adleman (RSA). But disadvantage of AES algorithm is sharing of key. There is no safe way to share the key. And there is also loss of data when we compresses large file. These algorithms had some security issue related with key length, block size, security rate and execution time [9].

AES Implementation and Compression: [6] To secure data while transmission on the network, it is must to encrypt the data and upload it in unreadable format. Compressing the data reduces the size of data and is required to save memory space and transmission time with security [12].In the process of compression, it removes extra space characters inserting simple repeat characters to indicate a string of repeated characters and substituting smaller bit strings for frequently occurring characters[7].
Encryption techniques used is symmetric encryption approach. In the proposed technique there is a common key between sender and receiver, which is known as private key. The private key concept is the symmetric key concepts where plain text is converted into encrypted text known as cipher text using private key where cipher text decrypted by same private key.

Features after Post Implementaion:

- Using AES encryption the size of the file increases as it does padding at the end of the file.
- The time taken by all different format of files or datasets is evaluated same, no matter if it is a text file, audio file or a video file.
- With GZIP compression technique the size of file at the time of upload will save space when initially encrypted and then compressed [12]. Hence GZIP compression technique will be used instead of LZ4 compressions.
- It is model proposed may found adaptable with different data sets e.g. audio,video,text etc. when implemented using parallel and distributed computing system i.e. Hadoop's Map Reduce. It can perform encryption in parallel where users can work automatically in parallel. In future I will implement this model and verify the assumption by evaluating the performance AES encryption algorithm with compression.

Conclusion

In this paper we have discussed about Hadoop technologies, its components, benefits of HDFS and Map Reduce. With the explosion of data, an oraganizations is shifting towards big-data management system. In this context, it is important to discuss about various technological challenges and its security issues. The proposed solution introduces one more layer as Security layer with proposal of AES implementation with compression in it.

References

[1] Zhao J., Wang L., Tao J.,Chen J., Sun W., Ranjan R., et al., "A security framework in G-Hadoop for big data computing across distributed Cloud data centres," Journal of Computer and System Sciences, vol. 80, pp.994-1007, 2014

[2] Kadre Viplove , Chaturvedi Sushil , "AES – MR: A Novel Encryption Scheme for securing Data in HDFS Environment using MapReduce ", http://www.ijcaonline.org/research/volume129/number12/kadre-2015-ijca-906994.pdf,International Journal of Computer Applications (0975 – 8887) Volume 129 – No.12, November2015.

[3] Gaikwad Rajesh Laxman , Prof. Dhananjay M Dakhane and Prof. Ravindra L Pardhi," Network Security Enhancement in Hadoop Clusters", http://ijaiem.org/Volume2Issue3/IJAIEM-2013-03-23-065.pdf, International Journal of Application or Innovation in Engineering & Management (IJAIEM) ,Volume 2, Issue 3, March 2013 ISSN 2319 – 4847.

[4] Saraladevia B.,Pazhanirajaa N., Victer Paula, Saleem Bashab, Dhavachelvanc P.," Big Data and Hadoop-A Study in Security Perspective", 2nd International Symposium on Big Data and Cloud Computing (ISBCC'15).

[5] Karthik D, Manjunath T N, Srinivas K," A View on Data Security System for Cloud on Hadoop Framework", http://research.ijcaonline.org/nckite2015/number3/nckite2661.pdf, International Journal of Computer Applications (0975 – 8887) National Conference on Knowledge, Innovation in Technology and Engineering (NCKITE 2015).

[6] Vinit G. Savant," Approaches to Solve Big Data Security Issues and Comparative Study of Cryptographic Algorithms for Data Encryption ",http://ijicar.com/wp-content/uploads/2015/04/RJ010106.pdf, Volume 1 : Issue 1 International Journal of Integrated Computer Applications & Research (ijicar) idin rJ010106 ISSN 2395-4310 2015 © IJICAR [http://ijicar.com].

[7] Monika Kumari ,Dr.Sanjay Tyagi ,"A Three Layered Security Model for Data Management in Hadoop Environment " https://www.ijarcsse.com/docs/papers/Volume_4/6_June2014/V4I6-0105.pdf , Volume 4, Issue 6, June 2014 ISSN: 2277 128X International Journal of Advanced Research in Computer Science and Software Engineering Research Paper Available online at: www.ijarcsse.com.

[8] B. Saraladevia, N. Pazhanirajaa, P. Victer Paula, M.S. Saleem Bashab, P. Dhavachelvanc ," Big Data and Hadoop-A Study in Security Perspective ",

http://www.sciencedirect.com/science/article/pii/S187705091500592X,2nd International Symposium on Big Data and Cloud Computing (ISBCC'15).

[9] Ms. Chetana Girish Gorakh, Dr. Kishor M. Dhole,"A Review on Security Approach in Big Data", http://www.iosrjournals.org/iosr-jce/papers/conf.15013/Volume%2010/9.%2037-40.pdf?id=7557 , IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p-ISSN: 2278-8727 PP 37-40 www.iosrjournals.org.

[10] Al-Janabi, Rasheed, M.A.-S., "Public-Key Cryptography Enabled Kerberos Authentication", IEEE, Developmeashents in E-systems Engineering (DeSE), 2011.

[11] Zvarevashe Kudakw, Mutandavari Mainford, Gotora Trust, , "A Survey of the Security Use Cases in Big Daa",http://www.ijircce.com/upload/2014/may/13_ASurvey.pdf, International Journal of Innovative Research in Computer and Communication Engineering,(An ISO 3297: 2007 Certified Organization),Vol. 2, Issue 5, May 2014

[12] Mehak, Gagandeep, "Improving Data Storage Security in Cloud using Hadoop", http://www.ijera.com/papers/Vol4_issue9/Version%203/U4903133138.pdf, Int. Journal of Engineering Research and Applications, www.ijera.com ISSN: 2248-9622, Vol. 4, Issue 9(Version 3), September 2014, pp.133-138

[13] Bhojwania Nikita, Prof. Vatsal Shahb," A Survey on HADOOP File System", http://ijiere.com/FinalPaper/FinalPaper2014112822174540.pdf, International Journal of Innovative and Emerging Research in Engineering Volume 1, Issue 1, 2014, e-ISSN: 2394 - 3343