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

Security of Big Data in Hadoop Using AES-MR with Auditing

Devika Tondon
M.Tech Student, Department of Computer Science and Engineering, SDDIET, Barwala, Panchkula, Haryana, India

Monika Khurana
Assistant Professor, Department of Computer Science and Engineering, SDDIET, Barwala, Panchkula, Haryana, India

DOI: 10.23956/ijarcsse/V6I5/0387

Abstract- While studying the various papers , I had gone through various methods and techniques adopted by author in securing the Big data.With the recent development in technology, networking and cost reduction in the storage devices, today we are flooded with the huge amount of data. Big Data analysis is now used in almost every state of our society, marketing, communication services, banking , research etc. The big data phase has shown the ways for many huge opportunities in the field of science, economic decision, educational system,healthcare system and novel forms of the public interaction and entertainment. But these opportunities also results into challenges in the area of privacy and security. Big data uses huge quantity of data that may be available in the cloud and it may require data processing distributed across the numerous servers. It is found that the development progress of big data also intensified and gave birth to the threats to the field of information security. This paper focuses on the big data and various privacy issues related to it.

Keywords Map reduce, Hadoop, HDFS, Kerberos, AES, VPN, AES-MR

I. INTRODUCTION

1.1 Big data

Big Data is the word used to describe massive volumes of structured and unstructured data that are so large and huge that it is very difficult to process this data using traditional databases and software technologies. The word Big data, is originated from the Web search companies who had to query loosely structured very large distributed data.[7] Big data is a current technology and also going to rule a world in future. It is the Buzz word hiding both the technical and marketing data inside it. The data that is small which collected in big size forms a terms called Big data and in real time.[4] The five main terms that signify Big Data have the following properties:[7]

a) **Volume**: Many factors contributes towards increasing Volume-storage transaction data, live streaming data and data collected from the sensors etc.
b) **Variety**: Today the data comes in all types of formats from traditional databases, text, documents, emails, videos, audio, transactions etc
c) **Velocity**: It means how fast the data is being produced and how fast the data needs to processed to meet the demand.
d) **Variability**: Along with the Velocity, the data flow can be highly inconsistent with the periodic peaks.
e) **Complexity**: Complexity of data also needs to be considered when the data is coming from the multiple sources. The data must be matched ,linked, cleansed and transformed into required and desired formats before actual processing

1.2 Big Data Issues

There are number of issues arising in big data. They are Management issues, Processing Issues, Security issues, and Storage issues . Each issue has its own task of surviving in big data and mainly focusing on the security issue.

a. Management Issues

The big data management is the collection of large volumes of Structured, unstructured and semi structured data from the organization, Government sector ,Private and Public Administration. The motive of the big data management is ensuring a high quality of data, data ownership, responsibilities, standardization, documentations and accessibility of data set. According to Gartner Big data Challenge Involve more than Just Managing the Volumes of Data .[4]

b. Storage Issues

The Storage is achieved using virtualization in big data where it holds the large set of Sensor information, media, video, E-business transaction records, Cell Phone Signal Coordinates. Many Big data Storage Companies Like EMC, IBM, Netapp, Amazon Handles a data in a Large volume by using some tools like NoSQL, Apache Drill, Horton Works, SAMOA, IKANOW, Hadoop, Map reduce, Grid Gain.

c. Processing Issue

The big data processing analyses the big data size in Petabytes, Exabyte or even in the Zetta byte either in Batch Processing or Stream Processing.

d. Security Issues

There are few challenges for managing a large data set in secure manner and inefficient tools, private and public database contain more threats and vulnerabilities, volunteered and unexpected leakage of data, and deficiency of Private and Public Policy makes a hackers to collect their resources whenever required. In Distributed programming frameworks, security issues start working when massive amount of private data stored in database which is not encrypted or in regular formats. Securing the data in presence of untrusted people is more difficult and especially when moving from the homogeneous data to the Heterogeneous data certain tools and technologies for massive data set is not often developed with the more security and policy certificates. Sometimes data hackers and system hackers involves in collecting a publicly available big data set, copy it and stores it in a devices like USB drives, hard disk or in Laptops. They involves in attacking data storage by sending some attacks like Denial of Service, Snoofing attack and Brute Force attack. If the unknown user knows about key value pairs of data it makes them to collect atleast some insufficient information. When the Storage of the data increases from single storage tier to Multi storage tier the security tier must also be increased. In order to reduce all these issues some cryptographic Framework techniques and robust algorithm must be developed in order to enhance the security of the data for future. Similarly some tools are developed like Hadoop; kerberos technology can be used for big data storage.[4].

1.3 HADOOP

Hadoop, which is a free, Java-based programming framework supports the processing of big sets of data in a distributed computing environment. It is a part of the Apache project sponsored by the Apache Software Foundation. Hadoop cluster uses a Master/Slave structure [6]. Using Hadoop , large data sets can be processed across the cluster of servers and applications can be run on systems with thousands of nodes involving the thousands of terabytes. Distributed file system in Hadoop helps in rapid data transfer rates and allows the system to continue with its normal operation even in the case of some node failures. This approach lowers the risk of the whole system failure, even in the case of a significant number of node failures. Hadoop enables a computing solution that is scalable, flexible, cost effective and fault tolerant. Hadoop Framework is used by popular companies like Google, Amazon, yahoo and IBM etc., to support their applications involving huge amounts of data. Hadoop has two mainly used sub projects – Map Reduce and Hadoop Distributed File System (HDFS).[7]

1.3.1 Map Reduce

Hadoop Map Reduce is a framework used to write the applications that process large amounts of data in parallel on clusters of commodity hardware resources in a very reliable, fault-tolerant manner. A Map Reduce job first divides the data into individual chunks which are then processed by Map jobs in parallel. The outputs of the maps sorted by the framework are then input to the reduce tasks. Generally input and output of the job are both stored in a file-system. Scheduling, Monitoring and re-executing of failed and not executed tasks are taken care of by the framework.[7].Hadoop‟s MapReduce seems to be an attractive cost effective solution for large scale data processing services like securing the data in cloud through block encryption.[6]

1.3.2 HDFS(Hadoop distributed file system)

The HDFS is the Java portable file system which is more scalable, reliable and distributed in the Hadoop framework environment. A Hadoop cluster contains the combination of a single Name node and group of the Data nodes. Using Commodity Hardware it provides redundant storage of large amounts of the data with the low latency where it performs the operations such like Write Once, Read Many Times. The files are stored with default size of 64MB as a block.The communication between the nodes occurs through Remote Procedure calls. It also decreases the data loss and prevents corruption of the file system.[4]. HDFS stores data on the compute nodes, providing high aggregate of bandwidth across the cluster. Both Map/Reduce and the distributed file system are designed in such a way that any node failures are automatically handled and thus provide us fault tolerance[8]

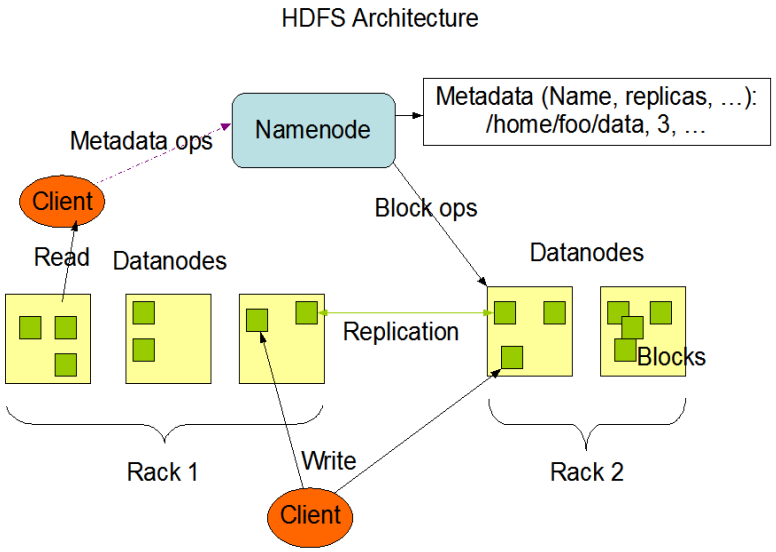

Fig 1 : Hadoop architecture

The Hadoop Distributed File System aims to emphasize on streaming data access, store large data sets and cope with hardware failure. HDFS follows the master/slave architecture and it has the namely Namenode which is master and DataNodes are slaves.

a) **NameNode (Master):** An HDFS consists of a single Namenode, a master server that manages the filesystem namespaces and regulates the access to files by clients. The Namenode makes the filesystem namespace operations like renaming,opening, closing, etc. of files and directories. It also determines the mapping of blocks to Datanodes.
b) **DataNode(slave):** This manages the storage attached to the nodes that they run on. The Data nodes are responsible for serving read and write requests from file system clients, also they performs the block creation, deletion, and replication upon instruction from the Namenode

II. HDFS SECURITY APPROACHES

The proposed work represented the different Approaches for securing data in Hadoop distributed file system .Each and every approach is based on different techniques and only has one aim i.e to secure big data.

2.1 Kerberos Mechanism

Kerberos is a network authentication protocol developed at MIT as part of the Project Athena. It uses private-key cryptography for providing the authentication across the open network and was developed before the popularity of the public key cryptography and systems like SSL. While many systems today use public key cryptography for the authentication, Kerberos also manages to do it with symmetric key cryptography In all, there are three steps that a client must take for access a service while using Kerberos, each of which involves a message exchange with a server:

1. ***Authentication*:**-A client must authenticates itself to the Authentication Server and receives the time stamped Ticket- Granting Ticket (TGT).
2. ***Authorization:*-**The client with the TGT, requests for a service ticket from the Ticket Granting Server(TGT).
3. ***Service Request:*-** The client uses the service ticket for authenticate itself to the server which is providing the service which client is using. In the case of Hadoop, this may or might be the name node or the job tracker. Together, the Authentication Server and Ticket Granting Server form the Key Distribution Center (KDC).[8]

2.2 Bull Eye Algorithm Approach

In big data, sensitive data can be credit card numbers, passwords, account numbers, personal details and many more are stored in a large technology called Hadoop. In order to maximizes the security in Hadoop, a new approach is introduced for securing the sensitive information which is called Bull Eye Approach algorithm. Also this approach is introduced on Hadoop module to view all the sensitive information in 360° to find whether all of the secured information are stored without any risk, and allows the authorised person to preserve personal information in a right way.[4]

2.3 The Advanced Encryption Standard (AES)

it is formal encryption method adopted by the National Institute of Standards and Technology of the US Govt., and it is accepted worldwide. The AES encryption algorithm is a block cipher that uses an encryption key and a many rounds of encryption. A block cipher is an encryption algorithm that works on a single block of data at one time. In the case of standard AES encryption the clock is 128 bits, 16 bytes in length. The word “rounds” is defined as the way in which the encryption algorithm mixes the data re-encrypting it to ten to fourteen times depending upon the length of the key. The AES algo itself is not a computer program or computer source code.[6]

2.4 Virtual private network

It is an open source solution that enables the implementation of virtual private network (VPN) for making secure point-to-point connections in routed or bridged configurations for secured access to remote machines. It makes use of the security protocol that employs SSL/TLS protocol suite for key exchange.It supports authentication amongst the peers by means of a pre-shared secret certificates, userid /password or key. When used in a multi client–server scenario, it allows server to issue or to give an authentication certificate for each client.[9]

2.5 AES-MR

An Advanced Encryption standard based encryption using MapReduce technique in MapReduce paradigm..The time taken for Performing the encryption and decryption process is relatively less for user generated content. Results show that AES-MR encryption process is found to be faster with mapper function alone in comparison with running the encryption process under mapper function and reducer function. Here a new encryption scheme is given by the combination of AES and MapReduce in order to secure data in HDFS environment.The results generated for encryption on different data proves that the proposed technique is well suited for protecting user generated sensitive data deployed in the HDFS environment.[1]

III. LITERATURE SURVEY

The several approaches available for implementing the Big data security in Hadoop are described as follows:

3.1 Kerberos:

An approach to implement Kerberos is laid by Rajesh Laxman Gaikwad and Prof. Dhananjay M Dakhane in 2013.They examined Hadoop clusters and security for hadoop cluster using Kerberos. The goal is to explore security problems Hadoop data users face with an advice on how to secure these environments.They explores the several methods through which a user can access data on hadoop clusters.

3.2 Bull eye approach algorithm:

This approach is given by B. Saraladevi and N. Pazhaniraja in 2015 for enhancing security in HDFS environment which can be achieved by this approach.Also their paper shows the Big data issues and focus more on security issues arises in Hadoop Architecture base layer called Hadoop Distributed File System (HDFS).

3.3 AES-MR(Advanced encryption standard using Map reduce)**:**

This technique is given by Viplove Kadre and Sushil Chaturvedi in 2015. They discussed the new technique to perform encryption in parallel. The time taken for Performing the encryption and decryption process is relatively less for user generated content. This technique is well suits for protecting user generated sensitive data deployed in the HDFS environment.

IV. METHODOLOGY

By observing the effectiveness of hadoop in different attack scenario. Hadoop consist of two core components: the job management framework that handles the map and reduce tasks and hadoop distributed file system. We introduce the syn flooding attack with the help of code attached to hadoop and then captured it with wireshark. Datanode of HDFS receives the blocks of data and deletes the flooded blocks and a fair scheduler for better job management in which small adhoc query jobs can be executed with periodic jobs (for monitoring) in parallel that prevents the degradation in performance of distributed file system.

Step1: Flooding on hadoop datanode

Step2: Capturing the live traffic

Step3: copying that file to hadoop user

Step4: job assignment

Step5: map and reduce task

Step6: delete the flooded blocks of data and auditing

V. PROPOSED WORK

5.1 Problem analysis

Encryption of large data stored in HDFS is actually a process which takes a lot of time and this time consuming nature of encryption should be controlled by encrypting the data using a parallel method. This study discusses a new technique to perform encryption in parallel using AES-MR (an Advanced Encryption standard based encryption using MapReduce) technique in MapReduce paradigm..The time taken for Performing the encryption and decryption process is relatively less for user generated content. Results show that AES-MR encryption process is found to be faster with mapper function alone in comparison with running the encryption process under mapper function and reducer function. This new encryption scheme is given by the combination of AES and MapReduce in order to secure data in HDFS environment .The results generated for encryption on different data proves that the proposed technique is well suited for protecting user generated sensitive data deployed in the HDFS environment. But Due to increased size of data files, size of log files (files which records the data activity), intrusion detection system faced so many problems and gives inaccurate results for detecting the attackers.

5.2Problem statement

To audit log files to find if any, malicious operations are performed or any malicious user is manipulating the data in the nodes. Also to solve the issue of data security at the storage level at HDFS by encrypting the data using AES-MR

5.3 Problem Formulation

We will focus on the type of packets which we are flooding in the hadoop framework. For this here an open source tools and software are used. We propose a technique which detect attack at hadoop datanode and analyze the working of hadoop distributed file system. We have shown hadoop's effectiveness in attack scenario, discussed various motivation for deployment.

VI. RESULTS

We introduce the attack packets like udp and ack with the help of code attached to hadoop and then captured these attack with wireshark. A code of java will be compiled in hadoop which will be used for encryption and auditing of files at datanode. Auditing is used to check whether the packet transferred is legitimate or not. On the basis of these types of results suspicious packets gets deleted and user who is sending those packets gets blocked which increases the level of security.

The following map and reduce graph show the map and reduce task .

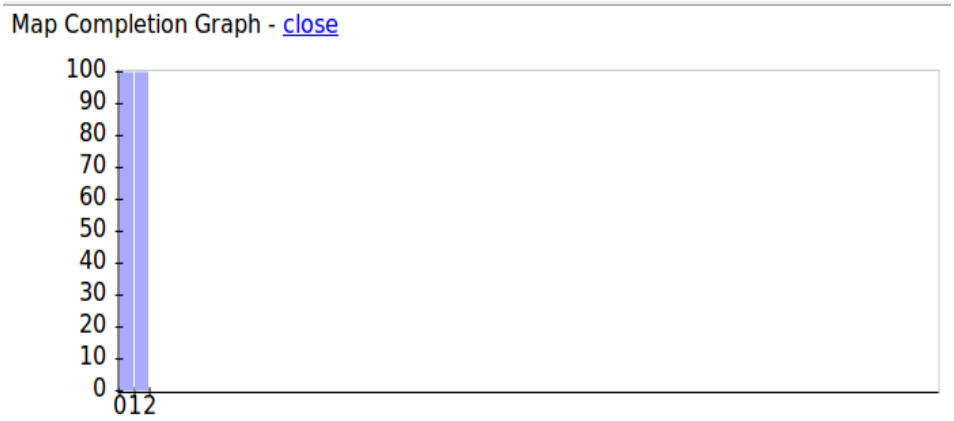

Fig 2: Map completion graph

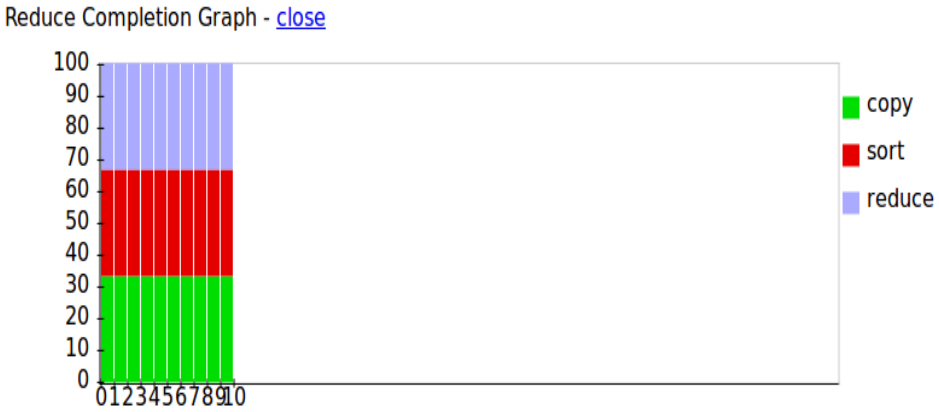

Fig 3: Reduce completion graph

VII. CONCLUSION

Flooding based attack involves large number of packets sent to the hadoop within a short span of time with the help of code attached to hadoop. We have justified it with wireshark. We use a distributed detection system to efficiently detect these attacks at an early stage. MapReduce technique of Hadoop is used for distributing the analysis task to idle workers in the Hadoop cluster and gets that job done efficiently and accurately. Datanode of HDFS receives the blocks of data and deletes the flooded blocks. Traditional scheduling methods perform very poorly in mapreduce due to two aspects: running computation where the data is and dependence between map and reduce task .

VIII. FUTURE WORK

We have implemented the detection of attack with a single node with its auditing. Our future work is to configure multiple nodes in hadoop and then introduce the dos attack. For that we need two ubuntu boxes. we will assign the IP address 192.168.0.1 to the master machine and 192.168.0.2 to the slave machine.

REFRENCES

[1] Viplove Kadre,Sushil Chaturvedi ,"AES – MR: A Novel Encryption Scheme for securing Data in HDFS Environment using MapReduce", International Journal of Computer Applications (0975 – 8887)Volume 129 – No.12, November 2015

[2] Daming Hu,Deyun Chen,Yuanxu Zhang and Shujun Pei,"Research on Hadoop Identity Authentication Based on Improved Kerberos Protocol",International Journal of Security and Its Applications Vol.9, No.11 (2015), pp.429-438 ,http://dx.doi.org/10.14257/ijsia.2015.9.11.39

[3] Vinod Sharma,Prof. N.K. Joshi,"The Evolution of Big Data Security through Hadoop Incremental Security Model",International Journal of Innovative Research in Science,Engineering and Technology Vol. 4, Issue 5, May 2015

[4] B.Saraladevi, N.Pazhaniraja, P.Victer Paul, M.S. Saleem Basha, P.Dhavachelvan,"Big Data and Hadoop-A Study in Security Perspective",Procedia Computer Science 50 (2015) 596 – 601,www.sciencedirect.com

[5] Kalyani Shirudkar,Dilip Motwani,"Big-Data Security",International Journal of Advanced Research in Computer Science and Software Engineering,Volume 5, Issue 3,March 2015

[6] Gazal ,Pankaj Deep Kaur,"A Survey on Big Data Storage Strategies",2015 IEEE

[7] Ashwin Kumar TK, Hong Liu, Johnson P Thomas, Goutam Mylavarapu,"Identifying Sensitive Data Items within Hadoop",2015 IEEE

[8] Devika Tandon,"A survey on security of big data in hadoop" ,International journal of research development and technology, Volume-5,Issue-6 (June-16) ISSN (O) :- 2349-3585

[9] Mehak,Gagandeep,"Improving Data Storage Security in Cloud using Hadoop",Mehak Int. Journal of Engineering Research and Applications,ISSN : 2248-9622, Vol. 4, Issue 9(Version 3), September 2014, pp.133-138

[10] Venkata Narasimha Inukollu ,Sailaja Arsi,Srinivasa Rao Ravuri,"SECURITY ISSUES ASSOCIATED WITH BIG DATA IN CLOUD COMPUTING"International Journal of Network Security & Its Applications (IJNSA), Vol.6, No.3, May 2014

[11] Rajesh Laxman Gaikwad,Prof. Dhananjay M Dakhane,Prof. Ravindra L Pardhi,"Network Security Enhancement in Hadoop Clusters",International Journal of Application or Innovation in Engineering & Management (IJAIEM),Volume 2, Issue 3, March 2013, ISSN 2319 - 4847

[12] Vikas Saxena, Shyam Kumar Doddavula,Akansha Jain,"Open Access Implementation of a secure genome sequence search platform on public cloud-leveraging open source solutions",Saxena et al. Journal of Cloud Computing: Advances, Systems and Applications 2012

