DOI: http://dx.doi.org/10.26483/ijarcs.v8i8.4855

ISSN No. 0976-5697

Volume 8, No. 8, September-October 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

REVIEW ON PRIVACY PRESERVING DATA ANALYTICS USING CRYPTOGRAPHIC TECHNIQUE FOR LARGE DATA SET

Yashi Gupta
[1]M.Tech. Scholar Department of Computer Science & Engineering Lakshmi Narain College Of Technology Bhopal(M.P.)

Dr. Vineet Richhariya
[2]Professor and Head, Department of Computer Science & Engineering Lakshmi Narain College Of Technology Bhopal (M.P.)

Abstract: Security is one of the major concerns in information technology industry. It becomes more crucial very it comes at end of data. Nowadays, data has become the centric point of all industry and all are moving around data only. Subsequently, trend of IT is also changed and they focused more on data value rather than services. Morally, large data volume is generated with different variety and high velocity. Due to public and open behaviors security has become one of the major concerns in hadoop framework. system. Hadoop is becoming so popular because of its big data storage. Hadoop system is designed without any security so security is the major concern in hadoop. Thus, in this paper we will be discussing about the core problem which is security in HDFS. As it works as the storage of large data so security is becoming the major weakness in hadoop development. Hadoop has file system to store data, Hadoop Distributed File System (HDFS) and Map Reduce are the file system of hadoop in which it stores large data. With the increase in popularity of hadoop, there is also a demand in trend for more and more security. Without any security model the sensitive data stored in hadoop is not secure. Also a new trend is rising for the encryption of stored data to prevent the confidentiality of data. Over the period of past few years an attempt have been made to achieve some level of security in Hadoop by using data encryption technique. In this paper all the attempts have been done to describe and prevent the data security in HDFS.

Keywords: HDFS, Security, Encryption technique, Big Data, Data Analytics

I. INTRODUCTION

Hadoop was developed by Doug Cutting and Mike Cafarella in the year 2006. Hadoop is a open source framework which computes and process large data sets in distributed environment. It is a java based programming which runs applications on system with thousands of hardware nodes interconnected. This approach reduces the disastrous failure in system and unexpected loss of data. If any number of nodes becomes nonfunctional then also it doesn't matter much because of thousands of interconnected node. It also deals with thousands of terabytes of data and rapid transfer of data among nodes. Hadoop handles big data processing task. Hadoop is rapidly enhancing and gaining popularity because of handling of large datasets and computing thousands of applications. Multiple techniques have been developed so that better performance can be achieved. These applications which process petabytes of data will lead to data transformation. Hadoop is like an open source project which provides with a framework to deal with large datasets and distributed batch processing. This all provides popularity to Hadoop but as it is designed without any security model, so one of the weaknesses of Hadoop is security.

A. Overview of HADOOP:

Hadoop is continuously updating and its latest versions Hadoop 2 emerges with the improvement in scheduling and resource management with introducing YARN. YARN stands for yet another Resource Negotiator. It is responsible for Resource Manager and Node Manager, resource manager manages the resources and also utilize them, deploy them and node manager manages data node and also report the status of data node to resource manager.[1]

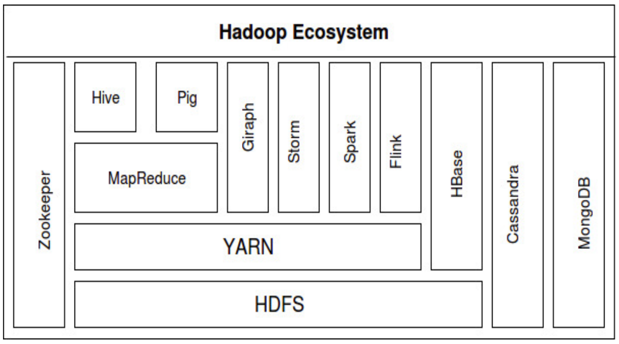

Figure 1. Hadoop Ecosystem

Apache Hadoop is also an open-source framework, used for processing large datasets and processing of distributed storage by using MapReduce programming. It consist of thousands of system and thousands of hardware, which are helpful in condition of failure or inoperative system. It is designed by keeping in mind that hardware failure are common and this failure should be handled automatically by framework. Hadoop as a software framework, composed of different components including Hadoop Distributed File System (HDFS), MapReduce Programming Model and Hadoop Kernel.[2][3] Hadoop divides the files into blocks and distributes them among nodes, it works as storage. Whereas, MapReduce Programming process the application stored in HDFS.[1]

B. Hadoop Distributed File System

Hadoop Distributed File System is a file system for Hadoop framework, it stores large files across cluster nodes. It replicates the data across multiple machines for the purpose of security and availability and therefore does not depends on RAID (Redundant Array of Independent Disk) storage but RAID configuration are used still to increase the performance of input output. Data is replicated and stored on three nodes with the default value of 3.[2]

HDFS is a portable and scalable file system it splits file into large blocks and distributes them to store on cluster node. Its replication factor replicates those file accordingly, this is the function of HDFS. But with all these features of HDFS, security is also essential.[4][6]

As HDFS manages complex application which is a challenging task, the huge data which is stored in HDFS are at risk because of missing of encryption at storage level.[5]

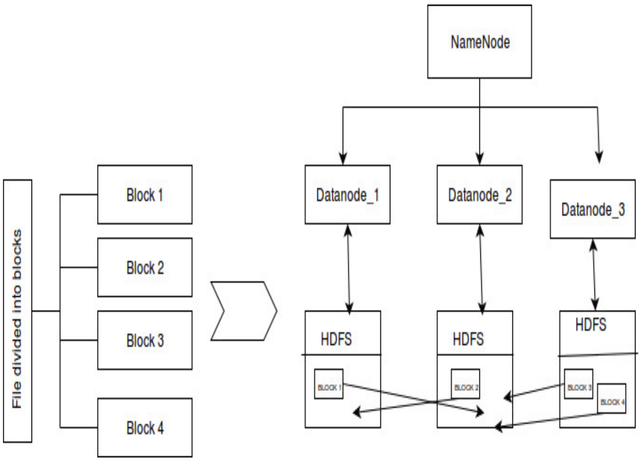

Figure 2. Working of HDFS

C. Security Issues in Hadoop:

The task of securing the whole data center is very essential and also achieving features like scalability, flexibility, performance and security challenges. Following are the security issues specified below:

1. Data access: Authentication and authorization plays an important role in security and limiting user access control of sensitive data.

2. Protection of data at rest: encryption is the protection of data at rest, which protects the access of data from outside. Encryption limits the replication of data.

3. Client interaction: Direct communication of client with data nodes and resource managers. Through it client can create integrity of data by sending malicious links or data[8][10]

.

II. LITERATURE SURVEY

Worked on Many technologies have been described and provided a better result regarding the security issue. With the enhancement in technology the security should also be increased.

Seon Young Park et al. In[1] with the increase in Hadoop system, security of stored sensitive data is also essential. To protect such data Hadoop file system requires the method of secure Hadoop.

Zerfos, Petros et al. In[2] described about some data protection techniques like Encryption. Throughout the life cycle of data its end-to-end security is required. HDFS security requirement and data protection can be achieved using some encryption techniques in Hadoop system.

Cheng, Zhonghan et al. In[3] explained that HDFS replicates the file and store that replicated file across cluster in multiple machine for the availability and durability. It is popular because of its scalable and distributed framework which allows big data applications to run.

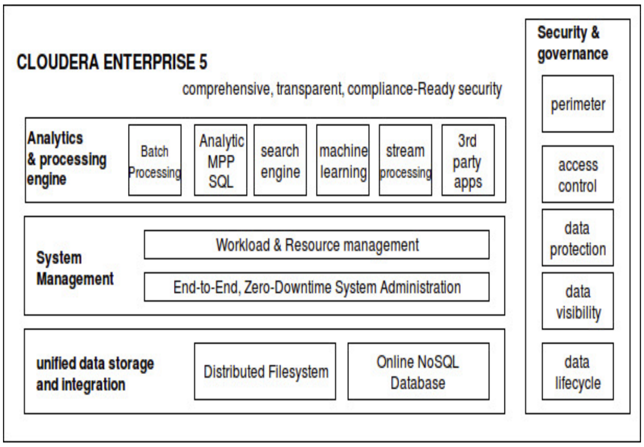

Figure 3. Future of Hadoop security

Shehzad Danish et al. In [4] abstracted about the initial phase of HDFS. When designing HDFS its initial phase i.e. storage efficiency and reliability is considered and data security was not considered. Due to improper data security data loss can be possible.

Transparent encryption in HDFS [5] author introduced about Transparent Encryption for Hadoop, which is an effort for data protection. Hadoop is an open source framework for cloud storage. It is a trending technology designed without security model for data storage, as it is a tool for storing large amount of data to increase the security of sensitive data and confidential information.

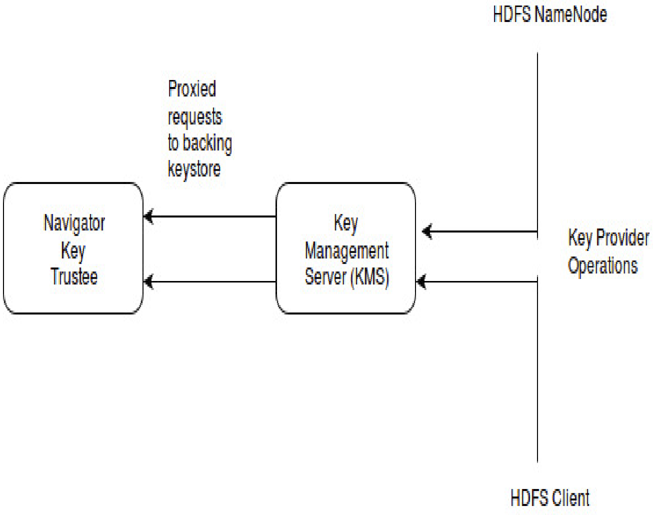

Figure 4. HDFS transparent encryption

Cloudera.com[6] Cloudera which is a commercial offering. “Security for Hadoop” is as essential and a important step in the originality of Hadoop and effort for data protection.

Owen O’Malley et al. In[7] introduces about the permission model implemented in HDFS, this model is for the files and directories. Author finds that if Kerberos is implemented over SSL then HDFS security can be enhanced by authentication and access control.

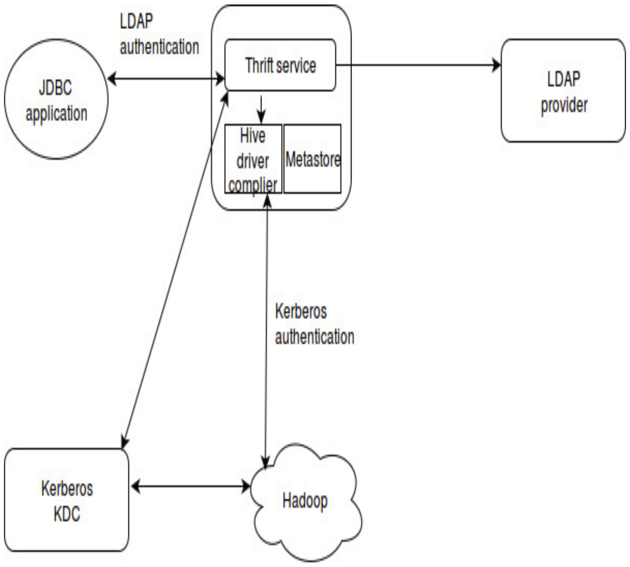

Figure 5. Security In Hadoop Using Kerberos

III. PROPOSED WORK

Relevant *Algorithm Used:*

The relevant algorithm used in our proposed work are :
1. Apriori Algorithm
2. RC6 Algorithm

A. Apriori Algorithm

In 1994, Agrawal and Srikant proposed Apriori algorithm. Apriori algorithm is an algorithm used for the mining of items that are used frequently and also for association rule learning. In a database every single item that appears frequently can extend as much larger till it appears in sets of items sufficiently. Association rule can be defined using Apriori algorithm because of frequency of appearing of items in datasets. [10]
Apriori Algorithm

It mainly operates for the frequently appearing items in a datasets and their collection.

1. Bottom-up approach is used in it.
2. It determines frequent sets of items.
3. Termination of algorithm can be done when no extension was found
4. Items can be said frequent if it appears at least 3 transactions in database.

Working steps of Apriori Algorithm :

1. Counts the appearance of items, occurrence of each items separately called Support (Value should be at least 3).

2. Generation of pairs of items appearing frequently in list form. Minimum support should be 3 of frequent item. [11]

B. RC6 Encryption

The RC6 stands for Rivest Cipher 6 which is a symmetric key block cipher. RC6 supports the key size up to 2040 bits and block size of 128 bits. It supports wide range of word length. RC6 is similar to RC5 and designed for the variety of word length. No key separation is required, it is flexible to all key size. It is derived from RC5.RC6 performs many operations like addition, subtraction, exclusive-or, multiplication, rotation to left-right.

RC6 performs encryption and decryption both.

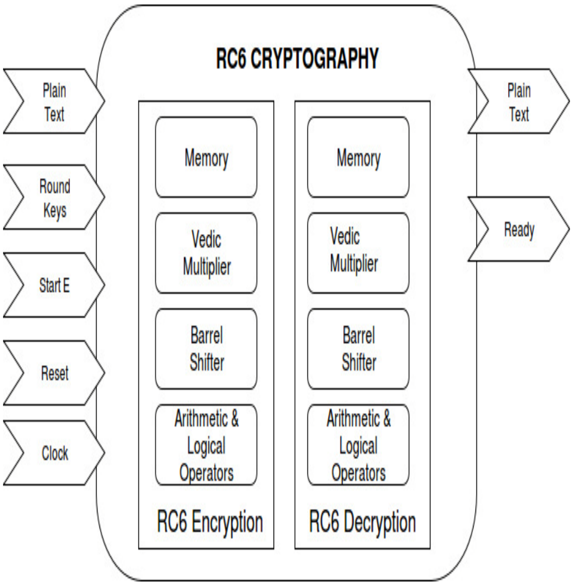

Figure 6. RC6 Block Diagram

RC6 provides with the advantage of security and high performance, fast and flexible and supports wide variety of word length.
Through the encryption flow diagram in RC6 algorithm we will be discussing about the technique of RC6 cryptography.

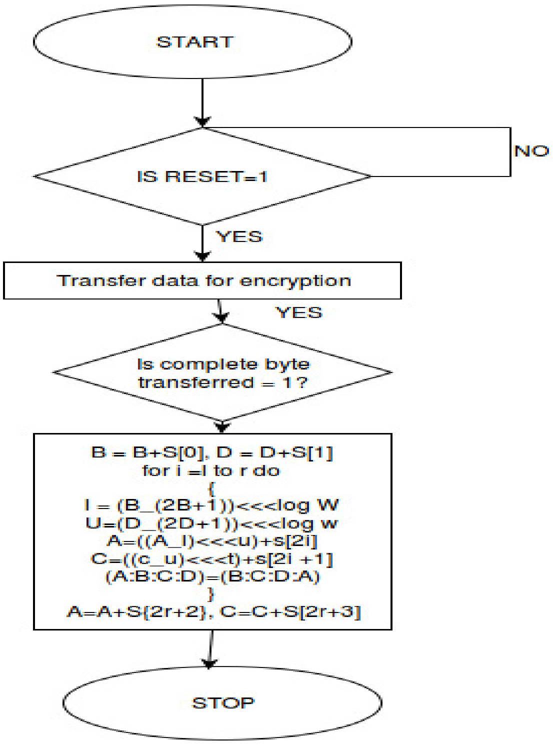

Figure 7. Flow chart for encryption in RC6

IV. CONCLUSION

The complete observation concluded that the created environment for storing data can be secured using encryption technique. Security is the major concern, and to achieve security encryption is done using RC6. In the complete work we have discussed about security which can be implemented using encryption technique where plain text is encrypted so that the stored data is in the form of encrypted data.

In the proposed system, RC6 encryption algorithm and Apriori algorithm are used. Apriori algorithm is used for the mining of items that are used frequently and also for association rule learning. RC6 algorithm is used for encrypting plain text so that its confidentiality cannot be stolen and misused. Apriori algorithm calculates confidence of support.

V. FUTURE WORK

With the need of security and implementation that has been proceeded in our work regarding the mitigation approach of security and its concern is a major deal. But still some implementation is required and should be worked on for testing on machines. Development is still required and needed and also processing so as to increase the speed of encryption on machine. Integration of key management system and increase in encryption algorithm are the future development.

VI. REFERENCES

[1] Seonyoung Park and Youngseok Lee, “Secure Hadoop with Encrypted HDFS”,Springer-Verlag Berlin Heidelberg in 2013.

[2] Zerfos, Petros, Hangu Yeo, Brent D. Paulovicks, and Vadim Sheinin. "SDFS: Secure distributed file system for data-at-rest security for Hadoop-as-a-service." In Big Data (Big Data), 2015 IEEE International Conference on, pp. 1262-1271. IEEE, 2015.

[3] Cheng, Zhonghan, Diming Zhang, Hao Huang, and Zhenjiang Qian. "Design and Implementation of Data Encryptionin Cloud based on HDFS." International Workshop on Cloud Computing and Information Security (CCIS 2013), pp. 274-277. 2013.

[4] Shehzad, Danish, Zakir Khan, Hasan Dag, and Zeki Bozkus. "A Novel Hybrid Encryption Scheme to Ensure Hadoop Based Cloud Data Security." International Journal of Computer Science and Information Security VOL 14, 2016 PP 480.

[5] Apache Hadoop "Transparent Encryption in HDFS." 2.7.2–.Accessed July 26, 2016.

[6] https://hadoop.apache.org/docs/r2.7.2/hadoop-project-dist/hadoop hdfs/TransparentEncryption.html.

[7] "HDFS Data At Rest Encryption". 2016. Cloudera.Com. Accessed July 26, 2016.

[8] https://www.cloudera.com/documentation/enterprise/5-4 x/topics/cdh_sg_hdfs_encryption.html.

[9] Owen O’Malley, Kan Zhang, Sanjay Radia, Ram Marti, and Christopher Harrell “Hadoop Security Design”, Technical Report, 2009.10

[10] S. Ghemawat, H. Gobioff, and S.-T. Leung, ``The Google _le system," in Proc. 19th ACM Symp. Oper. Syst. Principles (SOSP), 2003, pp. 29_43.

[11] S. Ghemawat and J. Dean, ``MapReduce: Simpli_ed data processing on large clusters," ACM Commun. Mag., vol. 51, no. 1, pp. 107_113,Jan. 2008.

[12] D. Borthakur, ``The Hadoop distributed _le system: Architecture and design," Hadoop Project Website, vol. 11, p. 21, Aug. 2007

[13] T. White, Hadoop: The De_nitive Guide. Farnham, U.K.:O'Reilly, 2012.