International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue XI, Nov 2018- Available at www.ijraset.com

Survey on Big Data Security based on HDFS using Clustering Technique

Divya B M[1], Vidyashree D. C[2], Rakshitha C. R[3], Shruthi H. H[4], Vanalakshmi A[5]
[1]*Assistant Professor, Department of Computer Science Engineering, BGSIT College of Engg.*
[2, 3, 4, 5]*Dept. of CSE, BGS Institute of Technology*

Abstract: ***Cloud computing stored for large data because of its ability to provide users with on-demand, reliable, flexible and low-cost services. Hadoop is most popularly used distributed programming framework for processing large amount of data with Hadoop distributed file system (HDFS) but processing personal or sensitive data on distributed environment demands secure computing. This paper presents result of K-Mediod algorithm, implemented on Hadoop Cluster by using Map-Reduce concept. In today's modern world, the availability of large number of online products like web-sites, web-portals, shopping sites, social networking sites, etc., give rise to a collection of extremely large and complex data sets known as Big Data. Proposed approach contains To address the data security issues in the Hadoop Distributed File System (HDFS).***
Keywords: Cloud storage, Hadoop, HDFS, Data Security, Encryption, Decryption and k-mediod clusters.

I. INTRODUCTION

Cloud computing is currently getting considerable attention in several communities, which provides the user's software resources, storage, and massive computing on demand [1]. Hadoop may be considered as a combination of Hadoop Distributed File System (HDFS) and MapReduce mode. HDFS reserves huge files (normally in the range of gigabytes to terabytes) over various machines. Hadoop is the platform using Distributed File System & capable of managing huge data-exhaustive applications and can run on commodity hardware. HDFS contains Namenode and Datanodes[2].

A. Big Data

Big data describes the large volume of data, it is a combination of huge datasets that can be handled using new techniques. Data that has extra-large Volume, comes from Variety of sources like text, audio, video, xml files etc., Variety of formats and comes at us with a great Velocity is normally referred to as Big Data. Big data can be structured, unstructured or semi-structured. Big data that hold the data generated by various equipment and applications like Black box. The term BIG DATA is simply used to describe the collection of complex and huge data sets such that it is difficult to analyse, store and process this kind of data using conventional database management tools and traditional databases management systems [3][4].

B. Big Data & Its Parameters

As the data is bigger from different sources in different form, it is represented by the 4 V's [3]:

1) *Volume:* Ratio of data or huge amount of data develops in every second. Machine develop data are examples for these components. Nowadays data volume is increasing very quickly from gigabytes to peta-bytes.
2) *Velocity:* Velocity is the different rates at which data is developing and processed. For example facebook, google, twitter etc
3) *Variety:* Variety is important characteristic of big data. It is a type of data. Data can be in different styles such as Text, numerical, xml files, application programs, images, audio, video data. On facebook more than 2 billion people are sharing files, photos, txt, videos, audio,
4) *Veracity:* Veracity means accuracy or anxiety of data. Data is uncertain due to the inconsistency and in completeness.

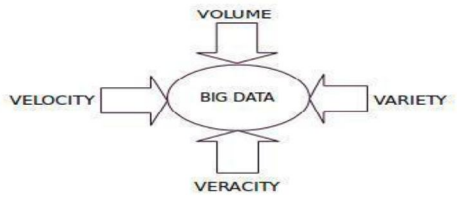

Fig. 1: Four – V's of BIG DATA

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue XI, Nov 2018- Available at www.ijraset.com

C. *Hadoop*

Two different technologies are merged to drive the Hadoop framework named as Map-Reduce and Google File System (GFS) [5]. Whereas Map-Reduce gives the state of art speedy technology to process huge amount of data, GFS provide the ability of automatic handling of node failure and provide the capability to framework to on heterogeneous commodity hardware. The first concept of Hadoop was published in 2004, but still very less support and documentation is available.

The Hadoop [6], is one of the recent trends in technology which is used as a framework for the cloud storage, is an open-source distributed computing framework implemented in Java and consists of two modules that are, MapReduce and Hadoop Distributed File System (HDFS)[4].

D. *Project Idea*

Hadoop is designed without considering security of data. Data stored at HDFS is in plaintext. This data is to be accessed by unauthorized user. So method for securing this data is needed.To reduce the uploading and downloading time of the files to and from the Hadoop Distributed File System (HDFS).Hence we are developing this highly secure system for Hadoop Distributed File System, using k-mediod clustering method for grouping the partial data.

E. *Need Of Project*

Hadoop is generally executing in big data clusters or might be in an open cloud administration. Amazon, Yahoo, Google, and so on are such open cloud where numerous clients can run their jobs utilizing Elastic MapReduce and distributed storage provided by Hadoop. It is a key to execute the security of client information in systems.

There are two types of cryptography key:

Secret Key Cryptography

Public Key Cryptography [8].

Secret key cryptography schemes are generally categorized as being either stream ciphers e.g. RC4 and OTP algorithm or block ciphers e.g. AES, DES, 3DES, and BLOWFISH Algorithms[11].

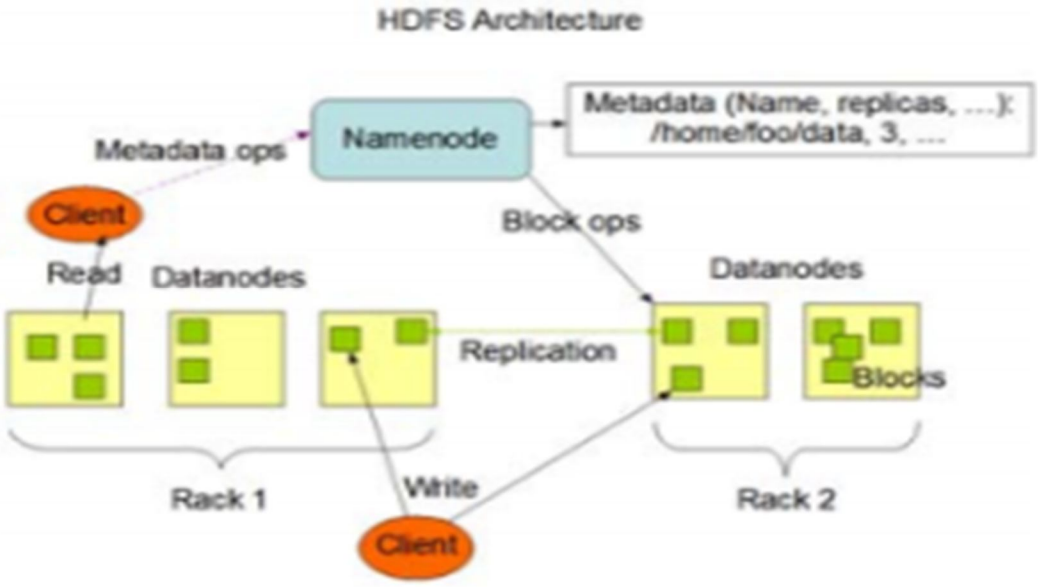

Fig: HDFS Arcitecture

II. RELATED WORK

Hadoop is a distributed file system which permits us to store enormous structured & unstructured information(i.e. Big Data). It is also helpful to process store huge amount of data in parallel environment. MapReduce is a powerful distributed processing model for large scale datasets. Hadoop is an open source framework and implementation of MapReduce. HDFS is designed mainly to handle big size files, so the processing of massive small files is a challenge in native HDFS. [7].

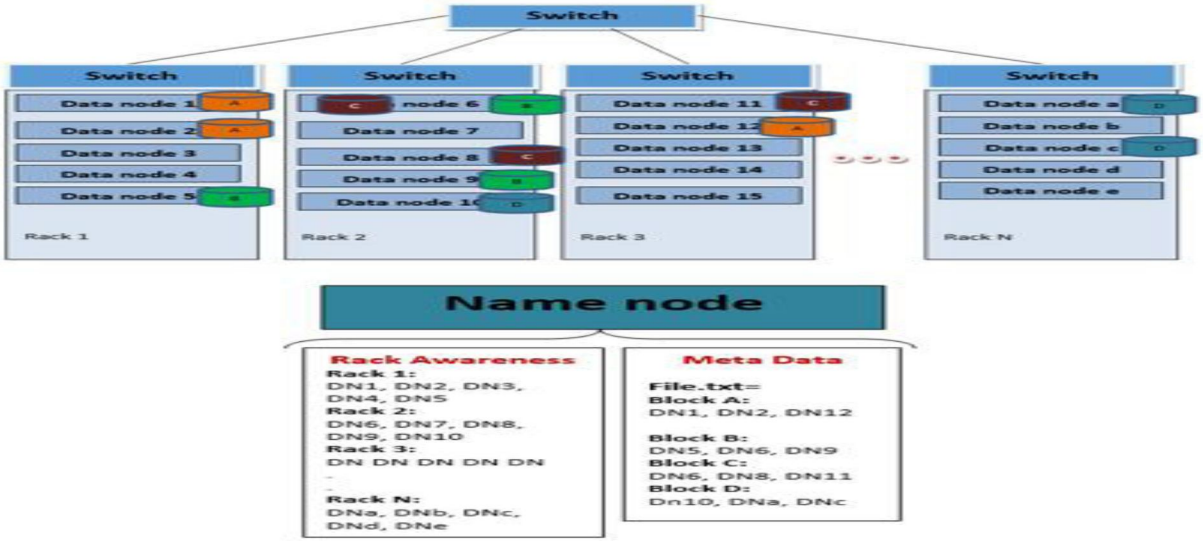

Fig 2: Hadoop cluster model

A. HDFS Structure Model

To start our cluster we need to run all the Hadoop daemon such as namenode, datanode, job tracker and a task tracker on the machine through commanding „start-dfs.sh" and „start-yarn.sh" on the command prompt.We design a new HDFS structure model which main idea is to merge the small files and write the small files at source direct into merged file and then duplicates the merged file through data pipeline. In relatively HDFS the data files are written to datanode directly without considering the size of data files. So we introduce a new HDFS structure model where the data files are gone through an intermediate step in which data files are first find out to be small files or large file in the HDFS. For merging process, a computational machine is needed. So we introduce a fast-node to do these merging files. This fast-node is highly configured machine that easily merge the small files in short time in the data node.

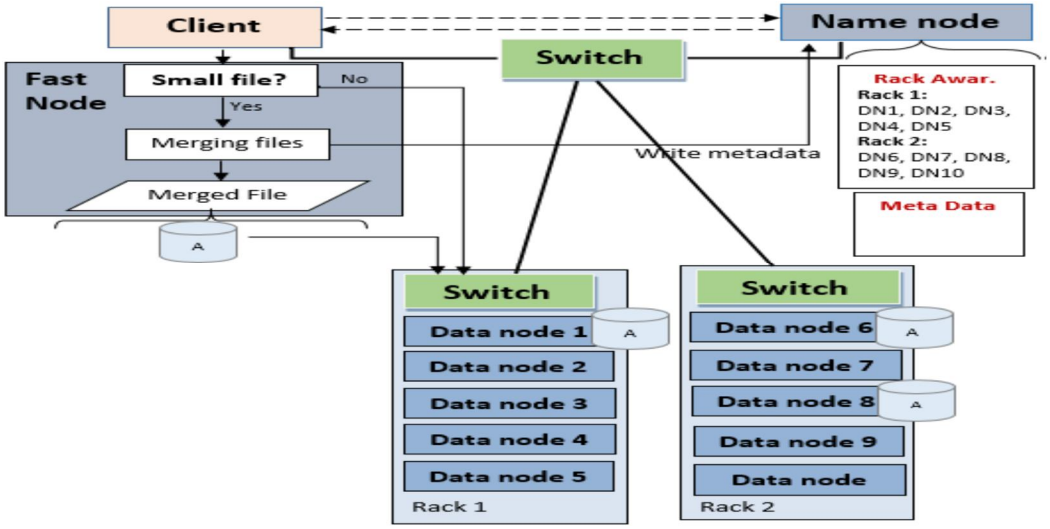

Fig 3: HDFS structure model

With the rapid growth of emerging applications like social network, semantic web,sensor networks and LBS (Location Based Service) applications, a variety of data to processed continues to witness a quick increase. Effective management and processing of large-scale data process an interesting but critical challenge. This paper introduces several big data processing techniques from system model and applications aspects. First, from the view of cloud data management and big data processing mechanisms, this paper contains key issues of big data processing, including definition of big data, big data management platform, big data service models, Hadoop distributed file system, data storage, data virtualization platform and distributed applications[9].

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue XI, Nov 2018- Available at www.ijraset.com

III. PROPOSED METHODOLOGY

A. Encrypting Files in HDFS

We assume that every file is encrypted before it is written to HDFS. Use the new technique encryption in Hadoop to encrypt the file, while buffering to HDFS and using file unstructured data. After encrypting the entire file, the HDFS starting to work with the encrypted file . These stages showed in "Fig."

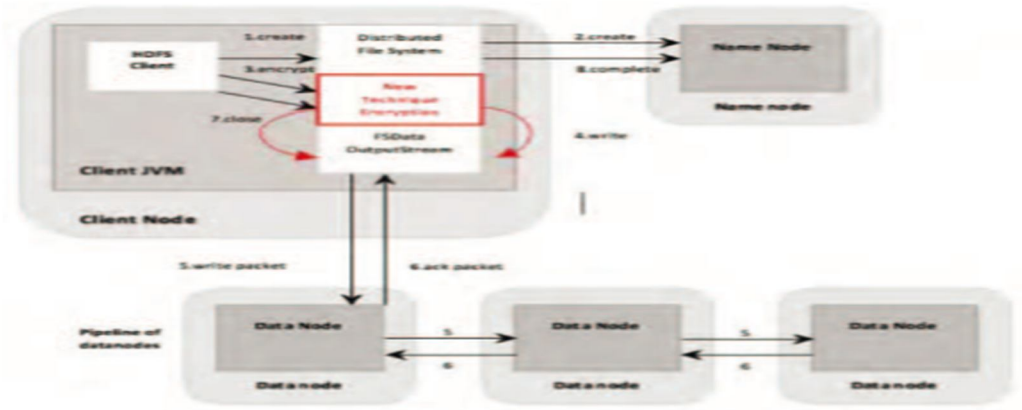

Fig 4: Proposed System: File Write Operation with Encryption

HDFS consists of a single master server named NameNode, which stores Metadata that manages the file system namespace and controls the access to the encrypted file used by clients. The encrypted file consists of one or more blocks that are stored in a set of DataNodes. The DataNodes are responsible for block creation, deletion, and replication upon instruction from the NameNode.

B. Decrypting Files in Map Task

The decryption process of our new technique will be executed parallel when MapReduce job started to read the data from HDFS blocks at the Datanodes, the decryption process at MapReduce shown in below figure. The decryption process of our new technique will be executed parallel when MapReduce job started to read the data from HDFS blocks at the Datanodes, the decryption process at MapReduce shown in "Fig". The MapReduce framework, which map is inputted to the reduction procedure performs the decrypted file. The MapReduce framework consists of one master JobTracker and one slave TaskTracker for each cluster node. JobTracker is responsible for scheduling any job component tasks on these slaves. Usually, it monitors and re-executes any failed task. The decryption process of our new technique will be executed parallel when MapReduce job started to read the data from HDFS blocks at the Datanodes, the decryption process at MapReduce shown in "Fig". The MapReduce framework, which map is inputted to the reduction procedure performs the decrypted file. The MapReduce framework consists of one master JobTracker and one slave TaskTracker for each cluster node. JobTracker is responsible for scheduling any job component tasks on these slaves. Usually, it monitors and re-executes any failed task.

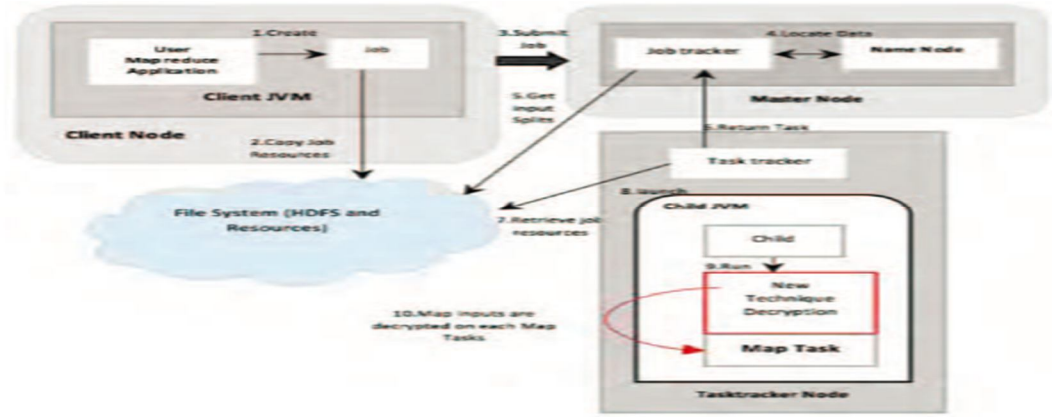

Fig 5 : Proposed System: File Read operation with decryption

The MapReduce framework, which map is inputted to the reduction procedure performs the decrypted file. The MapReduce framework consists of one master JobTracker and one slave TaskTracker for each cluster node. JobTracker is responsible for scheduling any job component tasks on these slaves. Usually, it monitors and re-executes any failed task[8].

C. Encryption/Decryption Mechanism

We will encrypt the file on the HDFS using a new approach, that consists of AES and OTP algorithms. We used AES algorithm in the cipher block chaining with the (ECB) mode which is one of the most popular block cipher algorithm and suitable for handling HDFS blocks, and we used OTP algorithm as a stream cipher to keep the plaintext in the same length, the file is symmetrically encrypted by big size of key length 384 bits, the key length of AES and OTP are set to 128 and 256 bits. The user keeps the private key in order to decrypt the file. In this case, we used the symmetrical encryption algorithm because it is safer and more suitable than asymmetrical encryption algorithm. When the user requested to upload the file to HDFS,the application server will generate the random key then divided it into two keys to using it for multi encryption and decryption algorithms AES and OTP.

IV. K-MEDIODS CLUSTER:

Clustering is a data mining method used to place data elements into related group. K-medoids clustering is a different of K-means that is more robust to noises and outliers. Instead of using the k-mean point as the center of a cluster, K-medoids uses an actual point in the cluster is represent it. k-Medoid is the most centrally located object of the cluster, with minimum sum of distances to other points. In contrast to the k-means algorithm, k-medoids chooses datapoints as center of the clusters (medoids or exemplars) and works with a generalization of the Manhattan Norm to define distance between datapoints of This method was proposed in the year of 1987[1] for the work with norm and other distances.k-medoid is a classical partitioning technique of clustering that clusters known A-priori. A useful tool for determining k is the silhouette. It is more robust to noise and outliers as compared to k-means because it minimizes a sum of pair wise dissimilarities or instead of a sum of squared Euclidean distances. A medoid can be defined as the object of a cluster whose average dissimilarity to all the objects in the cluster is minimal. i.e. it is a most centrally located point in the cluster[10].k-medoids is also known as partitioning method of clustering that alliance the data set of n objects into k clusters with k known a priori The k-mediod cluster is used for grouping partial data is offerd from the cloud services in the Hadoop distributed file system. we can use both Hadoop and the Map-reduce technique.

V. CONCLUSION

This paper address the data security protection in the cloud. To ensure data security in Hadoop data storage. In the proposed approach encryption, HDFS files are encrypted by using AES and OTP algorithms. To integrate the existing file write operation with the AES and OTP encryption algorithms. To integrate the existing file read operation with the AES and OTP decryption algorithms. To reduce the uploading and downloading time of the files to and from the Hadoop Distributed File System (HDFS).To reduce the size of the encrypted file to improve the storage efficiency and the performance of the HDFS.K-mediod performs well and the results totally depend on the size of Hadoop cluster. k-medoids have been performed well on the big data sets in the terms of elapsed time and clustering quality. k-medoids is better in all aspects of the performance analysis.

REFEERENCES

[1] S.Vikram phaneedra & E.Madhusudhan Reddy, “Big data-solutions for RDBMS- A survey” in 12th IEEE/IFIP Network operations & management symposium (NOMS 2010) (Osaka,Japan,Apr 19 2013).
[2] Dave Beulke, Big Data Impacts Data Management: The 5 Vs of Big Data, November 2011.
[3] “BIG DATA & HADOOP: A Survey “Rehana Hassan1, Rifat Manzoor2, Mir Shahnawaz Ahmad3
[4] S.Vikram phaneedra & E.Madhusudhan Reddy, “Big data-solutions for RDBMS- A survey” in 12th IEEE/IFIP Network operations & management symposium (NOMS 2010) (Osaka,Japan,Apr 19 2013).
[5] Apache Hadoop. http://hadoop.apache.org/
[6] ” An approach for Big Data Security based on Hadoop Distributed File system”Hadeer Mahmoud,Abdelfatah Hegazy,Mohamed H. Khafagy.
[7] M. B. Alam, “A New HDFS Structure Model to Evaluate the Performance of Word Count Application on Different File Size,” vol. 111, no. 3, pp. 1–4, 2015.
[8] An approach for Big Data Security based on Hadoop Distributed File system 2018 Hadeer Mahmoud,Abdelfatah Hegazy,Mohamed H. Khafagy
[9] Title: “Challenges Involved in Big Data Processing &,” vol. 5,no. Viii, pp. 841–844, 2017. D. Sharma, G. Pabby, and N. Kumar.
[10] Kaufman, L. and Rousseeuw, P.J. (1987), Clustering by means of Medoids, in Statistical Data Analysis Based on the –Norm and Related Methods, edited by Y. Dodge, North-Holland, 405–416.
[11] D. Elminaam, H. M. A. Kader, and M. M. Hadhoud, “Performance evaluation of symmetric encryption algorithms,” Int. J. Comput. Sci. Netw. Secur., vol. 8, no.12, pp. 280–286, 2008.