An Optimized Approach for Medical Image Retrieval and Storage in Hadoop Framework

Latika kakkar
Department of Computer Science Engineering
Chitkara University, India

Gaurav Mehta
Department of Computer Science Engineering
Chitkara University, India

Abstract- Security also has been major concern in every organization. Due to increase in data day by day, there are issues of storing data in such a way that it can be accessed and processed with more efficiency. Patient privacy and their medical records have become an important issue in hospital organizations. The inter-hospital sharing of medical records were based on papers and there were chances of duplicate prescriptions. To avoid duplicate prescriptions in medical data, electronic records are the possible solution. Electronic medical records are computerized records maintained by hospitals. By using electronic medical records patient privacy and health care efficiency can be improved which includes medical images and prescriptions. But inter-hospital sharing has been an issue. The motivation of this paper is to store and retrieve medical data and medical images between different hospitals more efficiently. The work has been implemented on cloud technology having Hadoop configuration. A medical image file accessing system is developed based on Hadoop in cloud. The experimental study showed improved efficiency and overall performance of the system.

Keywords – Big data, Cloud computing, Hadoop, Mapreduce, HDFS

I. INTRODUCTION

Due to increase in the data volume and variety from various sources it has become difficult to process and store data in efficient manner. 'Big Data' describes technologies and innovative techniques to manage and analyze petabytes or terabytes that can be structured, unstructured or semi-structured. For processing and analyzing huge amount of data in economical and competent manner parallelism technique is used. Big data has various characteristic including volume, variety and velocity. Volume defines the quantity of data which is increasing rapidly to terabytes and petabytes. Variety refers to various types of data like structured, unstructured and semi-structured data, images, audio and videos. Velocity defines the processing speed and time of data. There are various problems with processing of big data. These include heterogeneity and incompleteness, scalability, timeliness and human collaboration. Big Data is a data whose range, heterogeneity and elaboration require versatile architecture, algorithms, approaches to operate it and get important information from it. Hadoop is the major framework that can solve the problem of formulating Big Data, and solves the problem of making it useful for analytics purposes. Hadoop is an open source software project that enables the distributed processing of large data sets across clusters of commodity servers. It is designed to scale up from a single server to thousands of machines, with a very high degree of fault tolerance. Medical data and images have been so vast and have complexities that are difficult to address. The data sets of medical field are increasing and getting larger so it becomes difficult for traditional systems to process this medical data. This medical big data requires to be processed such that it can easily be accessed and used. Apache Hadoop Foundation provides a framework for storing large scale data on commodity clusters. This enables parallel fault-tolerant analysis of medical big data using the MapReduce [1, 2] programming model. In today's world, to utilize wealth of medical big data Hadoop based cloud data storage can be a great success. Medical records are very important documents that stores medical images, data and information of patient. These records are utilized by various medical institutions that might be working on team basis. The hospital stores data in computerized records as well as on paper. But there are various issues in this practice such as time constraint by human ingress and communication, problem in data backup, paper cost etc. An Electronic Medical Records (EMR) structure specifies that patient data is digitized, computerized and paperless. This method reduces documentation errors and improves productivity by integrating the processes. Patients' privacy and reduced paper usage and various medical problems can be improved using EMR. Financial management of medical institutions and efficiency improves

including decision support systems. The cross-organization sharing of resources can also be implemented with the use of EMR. The cost to manage storage is growing and storage per terabyte is reducing. The storage volumes is growing faster and hardware price are declining. Cloud computing is a technology that promises lower costs, availability, disaster recoverability and high scalability. The medical image retrieval and replication is major requirement as hospitals have to manage increased medical images and pictures. The picture archiving and communication systems (PACS) are incapable of providing efficient responses to queries [16]. These are inadequate to handle large queries and data retrieval with limited bandwidth. To improvise medical health facilities, an appropriate strategy is required to access images with high speed. To solve various issued faced for storing, exchanging and sharing medical images, a medical image files accessing system (MIFAS) is developed on hadoop platform [3]. This paper provides a cloud based Hadoop for MIFAS. MIFAS is accessed by users to store, retrieve and share medical images between different health organizations.

HADOOP: An efficient approach for processing of big data

Hadoop is a programming framework that is used to sustain the processing of outsized data sets in a distributed computing environment. Hadoop project stimulates the augmentation of open-source software and a framework is provided for the designing of extremely scalable distributed computing applications [20]. Hadoop is the top-leveled project in Apache Software Foundation and it supports the development of open source software. Hadoop compromises of two parts: HDFS and map reduce model. HDFS [20] is specially designed data set with cluster of commodity hardware with streaming access pattern. HDFS called as hadoop distributed file system has functionality of fault tolerance. The Hadoop Distributed File System (HDFS) saves outsized files on multiple machines. Thus reliability is achieved by replication of data across node diminishing the requirement of RAID storage. There are clusters of data nodes in hadoop that stores data using block protocol. To rebalance data on datanodes and to ensure replication, these are connected to each other. HDFS is designed in such a way that it can store enormous amount of data, has scalability feature and it can also handle failure of data storage infrastructure without loss of data. Clusters of machines are created and coordinated by hadoop with inexpensive computers. Even if there is any failure, Hadoop continues to process the clusters without any loss of data. Hadoop works without interruptions by passing work to the other machines in the cluster. HDFS store the data on clusters of datanodes by breaking files into blocks. HDFS then distributes the blocks to its various servers. Each file is stored on three different servers.

MapReduce Architecture

Mapreduce is the programming model of hadoop. This model process the data stored in HDFS by applying various operations on it. The programming model runs the data in parallel. In Hadoop, these kinds of operations are written in MapReduce jobs in Java. Map/reduce which performs operations on data two functions: map and reduce. Map function take input and provides intermediate key pairs. Reduce function merges all the values obtained in map function that have same intermediate key.

II. RELATED WORK

The benefits of medical images that are based on cloud were discussed in previous researches. These benefits included cost, time, scalability and duplication. Previous review revealed that a few medical images that are cloud based are implemented. A PACS system that was Hadoop based worked on medical images on cloud was also implemented. This system failed to provide an adequate interface. A Co-allocation mechanism has certain features which allow data downloaded from different nodes in parallel. There are lower faults in network and also co-allocation fastens the downloading of data. In [4] architecture was proposed that consist of an information service, local storage systems and a co-allocator [7]. The co-allocator searches the resources and finds the location of replicas from Information service [6].After that it provides the list of physical file.

Apache hadoop is a framework that processes data running on huge clusters of commodity hardware [8]. Reliability and movement of data is provided by Hadoop framework. A processing map/reduce model fragments the data into parts. Each fragment can be processed or re-processed on any cluster data node. Hadoop framework proves to be good replacement of PACS server as it provides a superior, valuable and scalable tool for health organization for image storage and retrieval. The similar work was accessible that defined an HPACS system. This system failed to present an appropriate management interface. Further studies showed the HDFS poor performance that causes issues in portability that includes file system allocation, disk scheduling under concurrent workloads and file system page cache overhead [9]. To improve the performance of HDFS application-level I/O scheduling can be used. Also by reducing fragmentation and overhead in cache Hdfs performance can be improved but portability factor has to be

ignored. But the portability of hadoop makes it easy for users by reducing the complexity of installation. Previous research [10, 17, 11, 12, 13, 15] shows that issue of data transfer in grid co-allocation can be can be unraveled. This concept is used in proposed system but in cloud environment. Google has developed various technologies like Google File System (GFS), MapReduce and Bigtable.GFS is which is also distributed file system has fault tolerance feature. MapReduce system and Hadoop Distributed File System (HDFS) are an open source utilization of the Google File System (GFS).

III. SYSTEM DESIGN AND IMPLEMENTATION

In the proposed work system design includes HDFS and a medical image file accessing system (MIFAS). HDFS of is used as distribution file system. MIFAS is a server that provides an interface for users. Through this interface queries for medical images can be executed. In proposed work, Google hadoop cloud services are used for the experimental study. Data is downloaded in parallel from data nodes of Google hadoop which improves downloading speed. User uses MIFAS to access medical images and other patient information. For efficient storage and retrieval of images and data, k-means clustering algorithm is used. When user uploads images through MIFAS the replication of medical images and data is done automatically using hadoop thus providing reliability and fault tolerance. In system workflow, the user can open portal and enter username and password for checking authentication. After successful authentication user can query patient details and medical images. Users enter an input value or keyword and can receive patient information. Google hadoop provides a secret key to establish connection to the datanodes. This key is used with java data object to access datanodes of hadoop. Storage and retrieval of data and medical images is practiced using the sub keys generated from this secret key. These sub keys are different for establishing different connections to hadoop. For login page two sub keys are used. For patient profile two sub keys are used. For doctor profile six keys are used for different connection.

SYSTEM WORKFLOW

1. Doctor/patient enters their username and password which will be checked at MIFAS server for authentication. If the user is authenticated then user will login successfully. There is also an option for registration for new user on the interface.

2. After login successfully, the user can upload the image for any patient. Images that are stored in data nodes are further encrypted using NTRU algorithm for security purpose. The Images are first converted into base64 then encryption is applied using NTRU algorithm. When the image upload button is clicked by user then he will receive one public key from MIFAS server. Then at client end it will generate one private key. After that whole image gets encrypted then MIFAS uploads that image on Google hadoop using java data object (JDO) API.

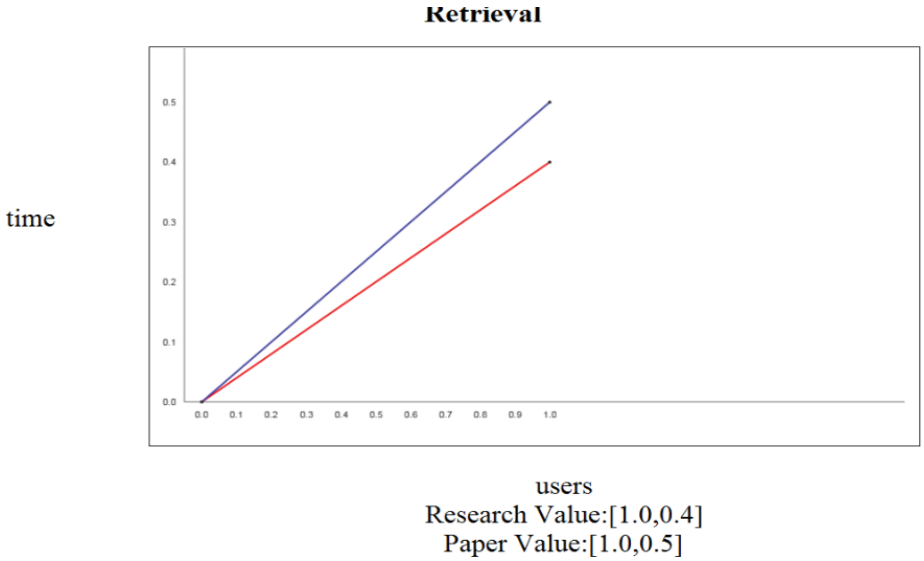
3. When the user will search data and images then request will first go to MIFAS server. After that MIFAS checks the query in their cluster array whether that file exist or not. If it exists then MIFAS provides that image to user from its own array. For this k-means clustering algorithm is used. If the image does not exist in MIFAS then it sends query to google hadoop using its API and fetch the result from there. The data retrieved from hadoop is stored in tree format using decision tree algorithm which improves the efficiency of result.

.

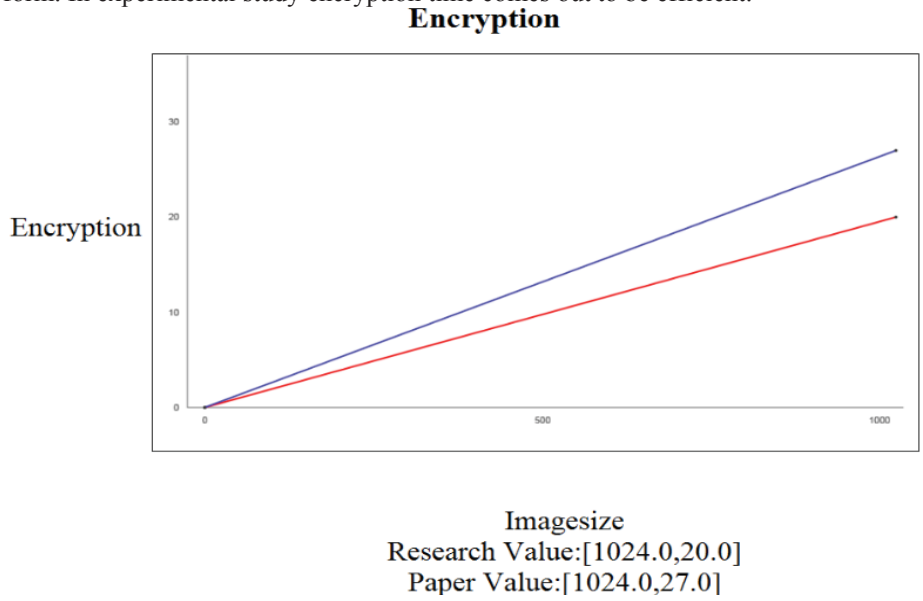
IV. EXPERIMENT AND RESULT

This section shows the experimental results based on the proposed algorithms. Four parameters are analyzed.

1. The experimental study analyzes the effectiveness of existing medical image files accessing system with the proposed system. File downloading speed is compared i.e. image retrieval time is compared which is less than the paper value.



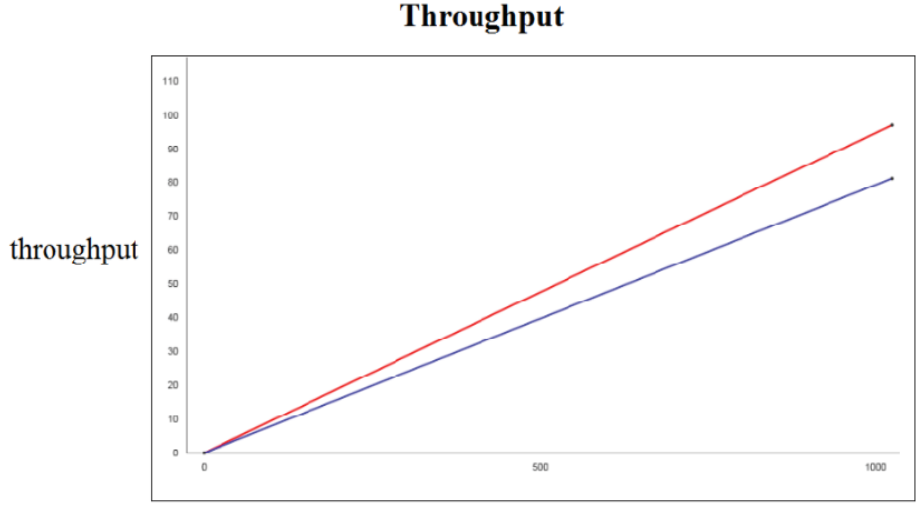
2. The time taken to encrypt medical data is analyzed and compared with previous encryption time. The time taken to encrypt data proves out to be less. Encryption is the process of conversion of plain text into cipher text or encrypted form. In experimental study encryption time comes out to be efficient.



3. The time taken to decrypt medical data is analyzed and compared with previous decryption time. The time taken to decrypt data proves out to be less. Decryption is the process of converting the encoded text into its original form which can be easily understood by user.

Imagesize
Research Value:[1024.0,20.0]
Paper Value:[1024.0,27.0]

4. Throughput of the proposed system is more efficient. Throughput is the measure of the amount of data processed within the given interval of time. It is the rate at which the provided data is processed. Less is the throughput more is the data processed.



imagesize
Research Value:[1024.0,97.0]
Paper Value:[1024.0,81.0]

V.CONCLUSION

In the proposed work a Medical Image File Accessing System (MIFAS) is developed which is based on HDFS of Hadoop in cloud. MIFAS enhances the functionality of storing, redeem and sharing of medical data and medical images. Medical images can easily be pooled between different hospitals. MIFAS thus obtained is scalable, cost-effective and provides replication and fault tolerance. The redundancy is obtained in medical images which is cost effective. File accessing performance is improved by improving the existing MIFAS. The proposed system provides

security by encrypting the images stored in datanodes of hadoop. As a result overall system performance is improvised. Future work can be to improve the system by enhancing the parameters. Also system can be implemented with the use of large number of hadoop distributed file system for storage purpose without hindering the system performance.

REFERENCES

[1] Dean J, Ghemawat S., "MapReduce: simplified data processing on large clusters", Communications of the ACM, 51(1), 107-13, 2008 Jan 1.
[2] Dean J, Ghemawat S., "MapReduce: a flexible data processing tool", Communications of the ACM, 53(1):72-7, 2010 Jan 1.
[3] Apache Hadoop Project. http://hadoop.apache.org/hdfs/
[4] Yang CT, Lin CH, Yang MF, Chiang WC, "A heuristic QoS measurement with domain-based network information model for grid computing environments", International Journal of Ad Hoc and Ubiquitous Computing, 5(4),235-43, 2010.
[5] Chervenak A, Foster I, Kesselman C, Salisbury C, Tuecke S., "The data grid: Towards an architecture for the distributed management and analysis of large scientific datasets", Journal of network and computer applications, 23(3), 187-200, 2000.
[6] Chervenak A, Deelman E, Foster I, Guy L, Hoschek W, Iamnitchi A, Kesselman C, Kunszt P, Ripeanu M, Schwartzkopf B, Stockinger H., "Giggle: a framework for constructing scalable replica location services", In Proceedings of the 2002 ACM/IEEE conference on Supercomputing , 1-17 , 2002.
[7] Chervenak A, Foster I, Kesselman C, Salisbury C, Tuecke S., "The data grid: Towards an architecture for the distributed management and analysis of large scientific datasets", Journal of network and computer applications, 23(3),187-200, 2000.
[8] Ganapathy G, Sagayaraj S., "Circumventing picture archiving and communication systems server with hadoop framework in health care services",Journal of Social Sciences , 6,310-4, 2010.
[9] Shafer J, Rixner S, Cox AL., "The hadoop distributed filesystem: Balancing portability and performance. InPerformance Analysis of Systems & Software (ISPASS)", IEEE International Symposium , 122-133, 2010.
[10] Yang CT, Fu CP, Hsu CH., "File replication, maintenance, and consistency management services in data grids", The Journal of Supercomputing, 53(3), 411-39, 2010.
[11] Chao-Tung Yang, Yao-Chun Chi, Ming-Feng Yang, Ching-Hsieh Hsu, " An anticipative recursively-adjusting mechanism for parallel file transfer in data grids", Concurr. Comput.: Pract. Exper, 22 (15), 2144–2169, 2010.
[12] C.T. Yang, I-Hsien Yang, Shih-Yu Wang, Ching-Hsien Hsu, Kuan-Ching Li, "A recursively-adjusting co-allocation scheme with a cyber-transformer in data grids" , Future Gener. Comput. Syst, 25 (7), 695–703, 2009.
[13] C.T. Yang, I.H. Yang, K.C. Li, S.Y. Wang, "Improvements on dynamic adjustment mechanism in co-allocation data grid environments", J. Supercomput, 40 (3), 269–280, 2007.
[14] Zhou Z, Chao SS, Lee J, Liu B, Documet J, Huang HK, "A data grid for imaging-based clinical trials", In Medical Imaging (pp. 65160U-65160U), International Society for Optics and Photonics, 2007.
[15] Yang CT, Wang SY, Chu WC., "Implementation of a dynamic adjustment strategy for parallel file transfer in co-allocation data grids", The Journal of Supercomputing, 54(2),180-205, 2010.
[16] Yang CT, Chen CH, Yang MF., "Implementation of a medical image file accessing system in co-allocation data grids", Future Generation Computer Systems, 26(8), 1127-40, 2010.
[17] Yang CT, Lin CH, Yang MF, Chiang WC, " A heuristic QoS measurement with domain-based network information model for grid computing environments" , International Journal of Ad Hoc and Ubiquitous Computing, 5(4), 235-43, 2010.
[18] Yang CT, Shih WC,Chen LT, Kuo CT, Jiang FC, Leu FY., "Accessing medical image file with co-allocation HDFS in cloud", Future Generation Computer Systems, 43, 61-73, 2015.
[19] Cohen JC, Acharya S, "Towards a trusted HDFS storage platform: Mitigating threats to Hadoop infrastructures using hardware-accelerated encryption with TPM-rooted key protection", Journal of Information Security and Applications, 19(3), 224-44, 2014.
[20] Shafer J, Rixner S, Cox AL, "The hadoop distributed filesystem: Balancing portability and performance" , In Performance Analysis of Systems & Software (ISPASS)" , IEEE International Symposium, 122-133, 2010.