ISSN No. 0976-5697

Volume 8, No. 3, March – April 2017

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

An Efficient Storage and Retrieval of DICOM Objects using Big Data Technologies

P. Haripriya
Department of Computer science,
Bharathiar University, Coimbatore, India

Dr. R. Porkodi
Department of Computer science,
Bharathiar University, Coimbatore, India

Abstract- DICOM is a global Information technology standard for electronic medical image. Meta data information and images are stored in a single file with dcm extension. A single DICOM file, sizes from MBs to GBs based on the study. PACS uses RDBMS to store and retrieve DICOM data. Replacing RDBMS with Big data technologies will help to handle DICOM data efficiently. All Indian government hospitals stores around 1400PB data collectively. Thus DICOM qualifies as a problem for handling big data. Storing and retrieving these big data from large repositories are highly complex and challenging. Applying big data techniques for handling DICOM data will helps to save many patient lives and improve areas like research, treatment methods, patient similarity searching, disease progression monitoring, clinical follow-up, case studies, training and learning, expertise sharing and helps to understand different patterns in the medical image data archive in a secured way. This paper presents an extensive survey on selective big data technologies for DICOM data storage and retrieval and also analyses the performance of Apache Pig, Hive and Spark while performing storage and retrieval of DICOM data.

Keywords*: Big data, RDBMS, PACS, DICOM, Storage*

I. INTRODUCTION

Digital image and communication in medicine [1] represents years of effort to create the most universal and fundamental standard in digital medical imaging developed by NEMA and AMR in 1993. As such it provides all the essential tools for diagnostically exact representation and processing of digital medical image. The DICOM is not just an image or file format. It contains all encompassing data transfer, storage and display protocol built and designed to wrap all functional aspects of medicine. [2] Dicom files use more than 2000 standardized attributes (data dictionary) to express various medical data such as patient information, study information, series information and image information. It also stores other kinds of information such as patient 3D position, size, radiation does, image processing filters, etc. Attribute with numerical and textual data types are candidates to be included in search and retrieval operations, being used as filters or as returning values for query expressions. All these information are stored in a single file with extension .dcm.

All the real world patient medical image data can be viewed by DICOM viewer software with relevant attributes. The DICOM viewer software shows the original dicom image and related all kinds of information. It is also playing series of images like video player. It is used for analyse the problem of patient scan image by radiologist.

Picture Archiving and Communication Systems (PACS) [1] is used to store DICOM file which use a Relational Database Management System (RDBMS) in the background. [3] The medical images are so far managed using RDBMS. Medical images are semi structured. The traditional Relational data model (RDBMS) handles structured information very well. RDBMS cannot handle semi-structured data efficiently. The characteristics of big data necessitate powerful and novel technologies to mine helpful information and permit more broad-based health-care solutions. Thus there is a need to look for an alternate solution to handle the medical images. The introduction of big data technology provides the different kinds of tools for handle the different type of structure data with more effectively in large repositories.

The management of medical data is an important domain of research. The Picture Archiving Systems (PACS) [4] systems are the currently used DICOM management system in most medical centres. These systems are very expensive and propose a low expressiveness (only per-defined queries over certain attributes). Additionally they do not cope with the heterogeneity problem, since they store mostly use a relational database that store all heterogeneous attributes in a blob-like data type without any ability to interrogate them. This paper provides an extensive study on various big data technologies so far used in the literatures for storage and retrieval and analysis of DICOM data. The research papers also investigated and compared the performance of big data technologies namely pig, hive and spark for storing and retrieving DICOM data of size MB to GB.

The paper is organized as follows Section II discusses the big data technology, section III Result and Discussion and finally the paper is concluded in section IV.

II. BIG DATA TECHNOLOGY

Big data [5] is a collection of huge amount of data (structured or semi-structured or unstructured) in the world repositories. Big data generates value from the store and processing of larger quantities of digital data information that cannot be analyzed with traditional computing techniques [6]. Big data is not just numbers, dates and strings. Big data is containing geospatial data, 3D data, medical image data, audio, video, unstructured text, including log files and social media. Traditional data base systems are used for smaller volumes of structured data, fewer updates or a predictable, consistent data structure. [7] Big data analysis includes different types of data. So the big data technologies overcome the traditional system limitation.

The Big data [8] characteristics are volume, variety and velocity. Big data implies enormous volumes of data.

The volume of data is generated by different resource systems like social media the volume of data to be analyzed is massive. Big data variety refers to the many sources and types of data both structured and unstructured. The data comes in the form of emails, photos, videos, monitoring devices, PDFs, audio, dicom file. Etc. This Big data variety of unstructured data creates problems for storage, mining and analyzing data. Big Data Velocity deals with the rapidity at which data flows in from different sources. The flow of data is massive and continuous, so it is very difficult to handle DICOM data.

The real time [9] big data is not just a process for storing petabytes or exabytes of data in a data warehouse, it's about the ability to make better decision and take meaningful actions at the right time. Fast forward to the technologies like Hadoop give you the scale and flexibility to store data before we know how we are going to process it. Big data Technologies such as Mapreduce, HDFS, Storm, Hive, Spark, heron and etc. enable you to run queries without changing the data structures underneath.

[11, 14] Roski and team presented the medical field contains all three of the 3Vs of big data; it has volume with approximately 500 petabytes of medical data generated through 2012 having a projected growth of 25,000 petabytes by 2020, it has variety with the various modalities (e.g. dictations, EEG, CT, MRI, DTI, PET, SPECT, MRS, fMRI, etc.) and it has velocity in the speed at which these modalities are created. Therefore, it is critical to use big data technologies to create new healthcare applications that are able to consume this vast and varying data.

Due to the characteristics of medical data applications [10] such as the heterogeneity, the extremely huge/ever-increasing size, and the expensive storage, it would be beneficial to exploit the power of cloud-based systems, like MapReduce , or its open source version Hadoop, Amazon SimpleDB, Amazon DynamoDB, Amazon RDS , SQL Azure, Pig, Hive to handle such challenges. This is because these systems provide promising solutions of cost-effectiveness, disaster recoverability, elasticity, Manageability and availability. Nevertheless, none of these systems consider the complexity of the DICOM format.

A. Apache Hadoop

Hadoop [11] is an open source frame work and Java-based programming framework that supports the processing and storage of extremely huge data sets in a different distributed computing environment. Abstract and facilitate the storage and processing of large and/or rapidly growing data sets, Highly scalability and availability, use commodity hardware with little redundancy, Fault-tolerance, Move computation rather than data.

Hadoop [12] is used for Data-intensive text processing, Assembly of large genomes, Graph mining, machine learning and data mining, large scale social network analysis. Failure of a single component must not cause the failure of the entire system only a degradation of the application performance, Failure should not result in the loss of any data. If any one component fails in this system, it must be capable to recover without restarting the entire system, Component failure or recovery during job must not affect the final output. The core of Apache Hadoop consists of a storage part, known as **Hadoop Distributed File System (HDFS),** and a processing part called **MapReduce**.

HDFS

Hadoop File System [13] was developed using distributed file system design. It is run on commodity hardware. Unlike other distributed systems, HDFS is highly fault tolerant and designed using low-cost hardware. It store huge data, the records are stored across several machines. These files are stored in redundant manner to liberate the system from probable data wounded in case of failure. HDFS also makes applications available to parallel processing. HDFS cluster primarily consists of a Name node and Data node.

Name Node that manages the file system Metadata .When we store a large file into HDFS, it is spitted into blocks (typically 128 MB, but user can override it). Then each block is stored into multiple Data Nodes independently. The name node acts as the master server and it does the many tasks such as Manages the file system namespace, Regulates client's access to files, it also executes file system operations such as renaming, closing, and opening files and directories.

Data Node that stores the actual data. These nodes handle the information storage of their system. The Data node performs read-write operations on the file systems, as per client request. They also perform operations such as block creation, deletion, and replication according to the instructions of the name node.

MapReduce

MapReduce [15] is a processing technique and a program model for distributed computing based on java. The MapReduce algorithm divided in to two important tasks, namely **Map** and **Reduce**. First task is Map takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs). Second task [25] is reducing task takes the output from a map task as an input and combines those data tuples into a smaller set of tuples. As the sequence of the name MapReduce implies, the reduce task is always performed after the map job task completed.

The major benefit of MapReduce is easy to scale data processing over multiple computing different cluster nodes. Under the MapReduce model, the data processing is called mappers and reducers. Decomposing a data processing function into mappers and reducers is occasionally nontrivial. But, once we inscribe an application in the MapReduce form, scaling the application to run over hundreds, thousands, or even tens of thousands of machines in a cluster is merely a configuration change.

B. Apache Hive

Hive [16] is a data warehouse infrastructure tool to process structured data in Hadoop. It provides a simple query these data by implementation Hadoop Map Reduce plans. It is a easy way to process large scale data, support SQL based queries, Provide more user defined interfaces to extend, programmability, efficient execution plans for perform, interoperability with other database. The Hive used for structured logs with rich data types (structs, lists and maps), user base wanting to access this data in the language of their choice, lot of traditional SQL workloads on this data (filters, joins and aggregations),other non SQL workloads.

C. Apache Pig

Apache Pig [17] provides an engine for executing data flows in parallel on Hadoop. Pig runs on Hadoop and

makes use of the Hadoop distributed file system (HDFS) and Map Reduce. Apache Pig [18] is an construct over MapReduce and it is used to analyze larger sets of data demonstrating them as data flows. It can make all the data manipulation operations in Hadoop using Apache Pig. To write big data analysis programs in Pig and it provides a high-level language known as Pig Latin. This language provides a variety of operators using which programmers can expand their own functions for reading, writing, and processing data. Apache Pig is generally used by data scientists for performing tasks. Apache Pig is used to process huge data sources such as medical resource, web logs, etc to perform data processing for search platforms, to process time sensitive data loads.

D. Apache Spark

Apache Spark [19] is a high-speed and general engine for big data processing, with built-in modules for streaming, SQL, machine learning and graph processing. Speed Run programs up to 100x faster than Hadoop MapReduce in memory, or 10x faster on disk. It cans admittance various data from different sources including HDFS, Cassandra, HBase, and S3. Spark was developed in response to limitations in the MapReduce cluster computing paradigm, which forces a particular linear dataflow structure on distributed programs. MapReduce programs contain different types of work such as read input data from disk or source, map a function transversely the data, reduce the results of the map, and store reduction results on disk. Spark's RDDs function as a working set for distributed programs that offers a restricted form of distributed shared memory.

Spark takes MapReduce [20, 21] to the next level with less expensive shuffles in the data processing with capabilities like in-memory data storage and near real-time processing. The performance can be several times faster than other big data technologies. Spark also supports lazy evaluation of big data queries, which helps with optimization of the steps in data processing workflows. Spark holds intermediate results in memory rather than writing them to disk which is very useful especially when we need to work on the same dataset multiple times. It is intended to be an execution engine that works both in-memory and on-disk. Spark operators perform external operations when data does not fit in memory. Spark can be used for processing datasets that larger than the comprehensive memory in a cluster. Spark will attempt to store as much as data in memory and then will spill to disk. It can store part of a data set in memory and the remaining data on the disk.

The extensive literature survey has been conducted in various big data technologies so far reported respect to DICOM data. The table 1 describe some of the productive research work conducted in the above field and the outcomes and demerits are analyzed.The outcome of the literature survey identified that Hadoop frame work is the best for storage and retrieval of DICOM images.

Table 1. Literature survey

Tool	*Author*	*Outcomes*	*Demerits*
MapReduce	Deligiannis P, Loidl H-W, Kouidi E, 2012, [22]	Predictive analysis time decreased from nine hours to only a few minutes when accessing a dataset of 10,000 real medical records.	Lacking in early systematic diagnoses.
Hadoop	Markonis D, Schaer R, Eggel I, 2012 [23]	Concurrent map tasks reduced the total runtime from 50 hours to 9 hours 15 minutes with maximum accuracy.	Computing efficiency and Increase response times at limited cost.
Hadoop	Gopinath Ganapathy and S. Sagayaraj, 2010, [24]	HPACS provided better result for storage, retrieve, performance and cost when compared with PACS.	Data replication
MapReduce	Jai-Andaloussi S, Elabdouli A, Chaffai A, 2013, [25]	Identified MapReduce task is effective mechanism for large database of CBIR and improves the efficiency significantly.	Not validated on larger image databases.
Hadoop-GIS,Hive	Fusheng Wang Ablimit Aji Qiaoling Liu Joel H. Saltz, 2013, [26]	Hadoop-GIS increased query efficiency, decreased loading time and support high performance with cost effective architecture.	Complex Query
Hadoop, HDFS	Xuguang Zhao, Shudong Zhang, Zhongshan Ren ,2015, [27]	Identified Hadoop HDFS is effective for large images and efficient file access and it reduces time for read and write.	Not suited for small size files.
Hadoop	YAO Qing-An, ZHENG Hong, XU Zhong-Yu, WU Qiong LI Zi-Wei 2 , and Yun Lifen, 2014 , [28]	Hadoop based medical image retrieval systems reduces the time of image storage and retrieval, and improve the image retrieval speed.	Low transmission speed of data and More time consumption.
Hadoop, HDFS	Chao-Tung Yanga,Wen-Chung Shih,Lung-Teng Chena, Cheng-Ta Kuoa,Fuu-Cheng Jiang a, Fang-Yie Leu, 2015, [29]	MIFAS is achieves acceptable redundancy in medical resources with much less expense, take less time, improve high reliability for data storage.	MIFAS does not suitable for small sized DICOM file.
Hadoop	Jyoti S. Patil* and G. Pradeepani, 2016, [30]	Hadoop based medical image retrieval system takes very less time for image retrieval and diseases are diagnosed automatically.	Less system performance
Hadoop	Sarmad Istephan, Mohammad-Reza Siadat , 2016, [31]	Hadoop framework is dramatic improvement in speed by store and retrieve, feasible and useful for medical queries.	Not suited for all medical images and more feature extraction required.
Hadoop, HDFS	R.KingsyGrace , R.Manimegalai , S. Suresh Kumar, 2014, [32]	Hadoop-CBIR approach takes less time with large number of images and facilitates accurate retrieval of images matching the queried image.	Data security and privacy to be strengthend.
Hadoop, HDFS	LI PJ , CHEN GJ , GUO WM ,2011, [33]	Hybrid storage architecture is suitable for storing and managing large volume of medical images using HDFS.	Data security, storage architecture and network topology issues to be strengthend.

III. RESULT AND DISCUSSION

This paper investigates the selective big data technologies for DICOM data storage and retrieval and to identify the better performing tools for DICOM data. So we compared Apache Pig, Hive, and Spark. DICOM files with different sizes (300MB, 900MB, and 1.2GB) are used to evaluate the performance of Apache Pig, Hive and Apache Spark. In Fig 1 and Fig 2 shows the performance of Apache Pig, Hive and Spark for DICOM store and retrieval. Apache Hive storage rate is 2.7s, 3.0s, 3.5s and retrieval rate is 40s, 52s, and 58s. Whereas Apache Pig storage rate is 0.70s, 0.42s, 0.38s and retrieval rate is 29s, 24s and 20s. This result clearly indicates that Apache Pig is better than Apache Hive for handling DICOM data. Apache Spark storage rates are 0.51s, 0.36s, 0.30sand retrieval rates are 27s, 21s and 17s. Thus for DICOM storage and retrieval, the Apache Spark is more efficient than Apache Pig and Hive.

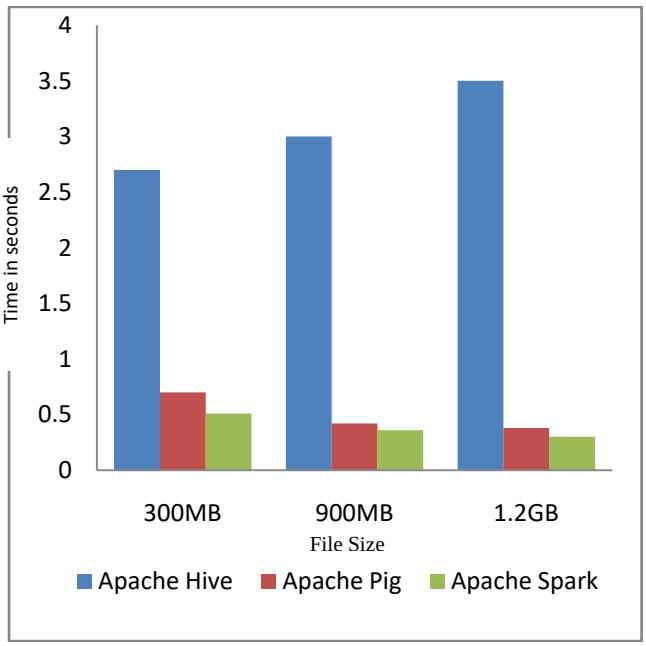

Fig.1 Store Data

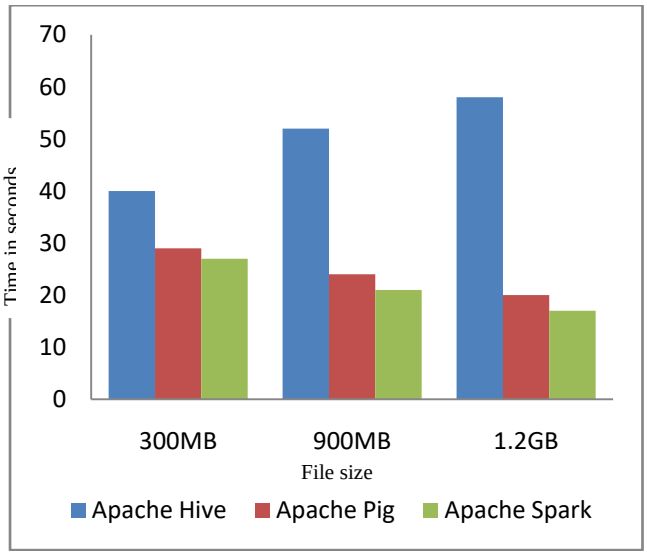

Fig. 2 Read data

This study also compares the performance of the mentioned tools for streamed input of DICOM data. Randomly selected DICOM files are sent using a simulator at the rate of 1.2 GB per minute. Apache Pig uses 30% of memory and Hive uses 45% of memory. Spark uses only 20% of memory. Spark betters other tools in CPU utilization also. Spark used only 5% of CPU. But Hive and Pig used 20% and 30% respectively. The performance of Spark is greatly improved when the data load is more. This study's result affirms that Spark is ideal solution for handling DICOM data. The results will vary based on system configuration.

The results conclude that Apache Pig performs better than Apache Hive. But Apache Spark outclasses both Apache Pig and Apache Hive for DICOM storage and retrieval. Thus Apache Spark is more suitable for handling DICOM storage and retrieval, than Apache Pig and Hive.

IV. CONCLUSION

Big data technology is an emerging approach for handling DICOM images. This study investigates several current big data solutions for efficient storage and retrieval for DICOM image and also analysed the performance of Apache Pig, Hive and Spark. The results show that, Apache Spark outclasses Apache Pig and Apache Hive for handling DICOM dataset. Apache Spark also supports distributed computing which helps to process the DICOM data in cloud environment. The performance of Apache Spark is improved as the data size is increased and also used less memory. Apache Spark can be a potential replacement for Hadoop in handing the DICOM data. Thus this study shows that Apache Spark can be used for effective storage and retrieval of DICOM object. But still, there are more areas to be explored on the basis of Meta data, scalability, security and my future research work will be on these topics.

V. REFERENCE

1. Digital Imaging and Communications in Medicine (DICOM): A Practical Introduction and Survival Guide by Pianykh, Oleg S.
2. Xiangrui Meng, Journal of Machine Learning Research 17 (2016) 1-7 Submitted 5/15; Published 4/16 MLlib: Machine Learning in Apache Spark,
3. A Performance Evaluation of Storage and Retrieval of DICOM Image Content in Oracle Database 11g Using HP Blade Servers and Intel Processors, An Oracle White Paper July 2008
4. Silva LA, Costa C, Oliveira JL. A PACS archive architecture supported on cloud services. *Int J Comput Assist Radiol Surg*. 2012;7(3):349–58.
5. Big data overview, Design of Enterprise System, University of pavia, 2013
6. BIG Data Analytics: A Framework for Unstructured Data Analysis T.K.Das1, P.Mohan Kumar2, International Journal of Engineering and Technology (IJET), 2013
7. Dilsizian SE, Siegel EL. Artificial intelligence in medicine and cardiac imaging: harnessing big data and advanced computing to provide personalized medical diagnosis and treatment. *Curr Cardiol Rep*. 2014.

8. Beyond Volume, Variety and Velocity is the issue of Big Data Veracity, by Kevin Normandeau, 2013
9. Ashwin Belle, Raghuram Thiagarajan, S. M. Reza Soroushmehr, Fatemeh Navidi,Daniel A. Beard, and Kayvan Najarian, Big Data Analytics in Healthcare, BioMed Research International Volume 2015 (2015), Article ID 370194, 16 pages
10. Big Data Application in Biomedical Research and Health Care: A Literature Review Jake Luo, Min Wu, Deepika Gopukumar, and Yiqing Zhao, Biomed Inform Insights. 2016; 8: 1–10. Published online 2016 Jan 19. doi: 10.4137/BII.S31559
11. H. Chang, Data-driven healthcare and analytics in a big data world, Healthcare Inform., (2015)
12. A. B. Mamulwar, T. Y. Wandile, S. A. Thorat, G. A. Bele, A Survey on Medical Image Retrieval Based on Hadoop , International Journal of Advanced Research in Computer Science and Software Engineering, 2015
13. Justin Kestelyn, Processing and Indexing Medical Images With Apache Hadoop and Apache Solr, 2014
14. J. Roski, G.W. Bo-Linn, T.A. Andrews, Creating value in health care through big data: opportunities and policy implications, Health Aff. 33 (7) (2014) 1115–1122.
15. Lotz, Marco, Extensible Distributed Processing Using A Cluster And Mapreduce With Applications On Lung Nodules Detection For Large Sets Of Tomographies, 2014
16. Hugo Pérez, Sergio Mendoza, Carlos Fenoy , Apache Hive, 2013
17. Apache Pig's Optimizer Alan F. Gates, Jianyong Dai, Thejas Nair Hortonworks
18. Programming Pig: Dataflow Scripting with Hadoop, By Alan Gates
19. New Horizons for a Data-Driven Economy, by José María Cavanillas, Edward Curry, Wolfgang Wahlster , 2016
20. Distributed Storage and Processing of Image Data , by Tobias Dahlberg , Linkopings University 2012
21. Big Data Analysis: Apache Spark Perspective By Abdul Ghaffar Shoro & Tariq Rahim Soomro
22. Deligiannis P, Loidl H-W, Kouidi E. Improving the diagnosis of mild hypertrophic cardiomyopathy with MapReduce. Proceedings of Third International Workshop on MapReduce and its Applications Date; Delft, Netherlands: ACM; 2012. pp. 41–8.
23. Markonis D, Schaer R, Eggel I, et al. Using MapReduce for large-scale medical image analysis. 2012 IEEE Second International Conference on Healthcare Informatics, Imaging and Systems Biology (HISB); La Jolla, California: IEEE; 2012.
24. Gopinath Ganapathy , Sagayaraj, Circumventing Picture Archiving and Communication Systems Server with Hadoop Framework in Health Care Services ,Journal of Social Sciences 6 (3): 310-314, 2010 ISSN 1549-3652 © 2010 Science
25. Jai-Andaloussi S, Elabdouli A, Chaffai A, et al. Medical content based image retrieval by using the Hadoop framework. In: 2013 20th International Conference on Telecommunications (ICT), IEEE. 2013
26. Wang F, Lee R, Liu Q, et al. Hadoop-GIS: A High Performance Query System for Analytical Medical Imaging with MapReduce: Technical Report; 2011; Atlanta: Emory University.
27. Xuguang Zhao, Shudong Zhang, Zhongshan Ren Implementation Based on Hadoop Ophthalmic Imaging Serialization File Store,2015
28. Yao, Zheng H, Xu, et al. Massive medical images retrieval system based on Hadoop. *J Multimed*. 2014;
29. Tung Yanga, Wen-Chung Shih b, Lung-Teng Chena, Cheng-Ta Kuoa, Fuu-Cheng Jiang a, Fang-Yie Leu a , Accessing medical image file with co-allocation HDFS in cloudChao, Future Generation Computer Systems,2015
30. , Jyoti S. Patil* and G. Pradeepani, Two Dimensional Medical Images Diagnosis using MapReduce, ndian Journal of Science an Technology, Vol 9(17), May 2016
31. Sarmad Istephan , Mohammad-Reza Siadat , Unstructured medical image query using big data – An epilepsy case Study, Journal of Biomedical Informatics 59 (2016) 218–226,2016
32. R. Kingsy Grace ; R. Manimegalai ; S. Suresh Kumar,Medical Image Retrieval System in Grid Using Hadoop Framework, Computational Science and Computational Intelligence (CSCI), 2014
33. LI PJ , CHEN GJ , GUO WM A distributed storage architecture for regional medical image sharing and cooperation based on HDFS , 2011

