Research on Medical Image Storage and Retrieval System Based on Hadoop

Xie Yaya¹, Zheng Bi-Geng²

¹School of Electronics and Information Engineering, Jingchu University of Technology, Jingmen 448000, China
²Periodical Press of Jingchu University of Technology, Jingmen 448000, China
33353302@qq.com

Abstract. With the Continuous Improvement of Medical Information, the Continuous Development of Image Processing Technology and the Continuous Upgrading of Hospital Image Acquisition Equipment, Medical Image Detection Has Become More and More Reliable and Effective, Becoming an Important Basis for Doctors' Diagnosis. Now Large Hospitals Use the Combination of Radiology Information Management System and Image Archiving and Communication System to Store and Access Patients' Images. with the Establishment of Regional Image Sharing and Cooperation Platform, the Online - Near Line - Offline Storage Strategy Adopted by the Traditional Image Archiving and Communication System Can No Longer Meet the Needs of Mass Image Data Storage. in the Retrieval of Mass Image Data, the Relational Database Used by the Image Archiving and Communication System is Inefficient in Reading the Image on the Scale of Mass Data, and the Existing Image the Archive and Communication System Lacks Complete Data Backup and Disaster Recovery Scheme. This Paper Analyses the Feasibility, Advantages and Problems to Be Solved of Using Hdfs to Store Massive Medical Image Files, and Studies How to Realize the Function of Precise Access and Range Access Provided by Pacs Efficiently in Hadoop.

1. Introduction
With the popularization of information technology and Internet technology, people have entered the era of big data. According to the statistics and analysis of existing data, it is expected that in 2020, the medical data will grow rapidly to 35zb (1zb = 230tb), which is 44 times of the data amount in 2009. How to manage and store these massive medical data and cumbersome data types reasonably has brought huge pressure to the medical industry. At present, the proportion of medical data in traditional Chinese medicine image data is more than 90%, which is a very important part of medical data. It has the characteristics of special file format, mostly small files (between 2KB and 1MB), large amount of data, fast growth rate, long storage time, etc. Moreover, medical image technology has made rapid development in the past ten years, with new technology and new equipment emerging constantly, and medical data is a very important part after being digitized and digitized, the medical image information presents a large amount of medical big data with rich diversity and huge storage, as shown in Figure 1. According to the survey, domestic hospitals generally adopt the three-level storage mode of online - near line and offline, the storage mode of disk array and tape drive with commercial centralized storage as the core, and the three storage architectures of NAS, DAS and San, which belong to the centralized file system of private network[1]. Although it has the advantages of high speed and good
isolation, with the passage of time, the data with the increase of quantity, the expansion performance is poor, the cost of hardware equipment is also very high, and it is difficult to achieve the sharing within the hospital or even in the region, so the flexibility is very poor.

Figure 1 Traditional Medical Image Storage

The emergence of cloud computing technology provides a new and effective way to deal with massive medical data. In addition to its significant advantages of resource integration, high availability, high performance and easy expansion, it provides a new method for data storage, retrieval, processing and analysis, which is very suitable for long-term storage and fast and effective access of medical image data[2]. At the same time, in view of the impact of small medical image files on the I/O efficiency of the server and network access, considering the use of distributed file system in different storage nodes to achieve data sharing, and provide multiple access nodes to improve performance, from the realization of a high-efficiency, strong scalability, high data security, and low cost storage scheme. Based on the above analysis, by making full use of the advantages of distributed file system and the excellent scalability of cloud platform, the performance research and optimization of mass medical image data storage are realized, which is of great research significance and application value.

2. Related Works

Since the emergence of medical image archiving and transmission system, the early traditional storage mode in the form of film has changed into digital storage and management mode. The promotion of reducing film to the point of no film greatly saves the cost of film, developer, film printer and image file storage space, which can save the hospital up to tens of millions of costs in a year. The research and development of PACS technology in China has entered a substantial stage in the past ten years, but there are only a few hospitals that can truly achieve image information sharing and remote real-time online diagnosis, especially the problem of mass data security storage, which has been the bottleneck of the rapid development of PACS. Some hospitals have purchased some foreign storage equipment and storage management software, but the one-sided pursuit of large and complete software results in waste of function and high price, which limits its wide application. In the process of implementing PACS, many technical problems need to be solved, the most important of which are transmission, display and image storage. At present, the main network system of each hospital is optical fiber. In addition, the price of workstation with high-resolution display is decreasing gradually, so transmission and display are not a problem. The requirement of image information archiving and storage is to meet a large number of data fast storage and easy to call. Therefore, it is very important to choose a suitable storage scheme. In 2006, Google first proposed cloud computing technology, which immediately ushered in the wave of world cloud technology. Many international companies began to join in the research of this technology, such as Amazon, Microsoft and other international multinational companies[3]. In cloud computing technology, open platform is the essence of cloud computing technology. Ordinary users can write programs through it like experts. The open source characteristics of cloud computing technology software enable users to obtain and improve code, which is also the reason why cloud computing can develop rapidly. Among them, cloud computing technology includes distributed computing processing technology, parallel computing processing...
technology and grid computing processing technology. These technologies are leading the development trend of the Internet. The development of cloud computing technology is also immeasurable for the good promotion of the medical industry[4].

Throughout the history of computer development, the earliest distributed storage technology is parallel file storage technology. At present, most of the research on medical image cloud storage uses HDFS as the underlying distributed file system. It supports offline bulk data processing, high throughput, and data block storage, which can save multiple copies, and easy to achieve load balancing. However, HDFS is suitable for small file data processing, and the smallest block is 64M, which is not suitable for medical image small file If HDFS is improved for small files, the engineering cost will be increased. Using lightweight distributed file system is a trend of mass medical image data storage, which is a very promising research direction.

3. Related Technologies of Hadoop Platform

Hadoop is a framework architecture that can store and process massive data in a distributed way. It uses map in map / reduce to divide the data into small pieces for processing, and then recombines them to realize the high-efficiency operation function. HDFS is deployed on a cheap PC, and data copies are stored repeatedly on multiple machines to achieve high reliability and fault tolerance of Hadoop. Hadoop can store and process data on a large number of clusters, so it has strong scalability. It can expand more nodes on the existing clusters for use. Hadoop supports Java programming framework, which can be implemented on cheap PC through JVM virtual machine. The main core of Hadoop is HDFS and Map Reduce. It is equivalent to map / reduce and Google File System (GFS) developed by Google lab[5, 6]. Through the characteristics of Map Reduce computing, it can process and compute large-scale data in parallel. Through the characteristics of HDFS distributed file system, it has a good fault-tolerant mechanism, low-cost requirements to realize its functions on the cheap cluster, and the advantages of ultra-high throughput. Of course, in addition to the two core institutions of HDFS and Map Reduce, it also includes a group of related projects. It includes a set of common components and interfaces of distributed file system and general I/O, Avro which can support serialization format, Map Reduce which is the framework of distributed data calculation, HDFS which is used to store data by Hadoop, pig which is the big data analysis platform running on Map Reduce and HDFS, hive which is proposed as a distributed and column storage data warehouse For SQL query on HDFS, running on HDFS similar to Google Big Table’s distributed NoSQL database for data transmission on database and HDFS, and zookeeper for distributed coordination service[7]. As shown in Figure 2, in the Hadoop architecture ecosystem, common provides common tools to support at the bottom, providing basic services and required APIs on the machine.

![Figure 2 Technologies of Hadoop Platform](image-url)
4. Overall Design of Medical Image Storage Architecture and Retrieval Application

In the case of large amount of medical image data, the data on the PACS system in the region is designed in a unified way, but it cannot meet the requirements of medical image data sharing. Through the analysis of the second chapter, the characteristics of HDFS can meet the requirements of medical image data storage and rapid processing, but it is not feasible to directly store DICOM3.0 standard image on HDFS, through the following discussion and analysis. HDFS and PACS systems have their own characteristics. HDFS can store Pb level massive data. When PACS is designed, it is limited to storing TB to hundreds of TB of data. HDFS also has the characteristics of low price, high reliability through data block copy, computing power in each storage node, and high throughput. Compared with PACS system, HDFS has its own defects. It is not a small file storage and cannot be applied in real time. Hadoop is a platform for storing and processing massive data under the current big data. As the core storage system of Hadoop, HDFS can satisfy all kinds of Internet big data processing companies such as Yahoo, Taobao and Baidu. Its big data storage characteristics can also satisfy the storage of medical image data. Combined with the actual application scenario, the paper builds the overall medical image storage architecture by combining the open-source his system, PACS system and distributed file system (introduced in Chapter 2). Combined with the application in the standard his system, it realizes the interface between the key his system, PACS system and distributed file system, and designs the “local cloud” two-level storage mode so that users can access his system directly access the local or cloud image files, and increase the cache to improve the overall access performance. The following will give a detailed introduction to each system and the overall architecture design. According to the long-term research and actual investigation, the use and test of two foreign open-source medical information management software, combined with the current situation of domestic medical image system, the design of Hadoop based medical image storage architecture is shown in Figure 3.

![Figure 3 Medical Image Storage Architecture](image_url)

According to the traditional mode used to manage various information data of hospitals in China, this architecture is composed of application layer, storage layer and platform layer, application layer is composed of clients of his system and PACS system, which is responsible for providing users with operation interface, information management, image viewing and other functions; storage layer is two-level storage mode of local end and cloud end, and local end is composed of his server and PACS server group which can be built on the local server, is responsible for storing and managing the hospital's structured information data and recent image data. The cloud is built by fast DFS large-scale distributed cluster, which is responsible for the permanent storage of long-term files. The data in the storage layer is uploaded by users and regularly uploaded from the local end to the cloud. The platform layer is built on the infrastructure by virtualization technology through the reasonable and effective use of server resources, it is convenient to provide cloud services. Most of the open-source software selected in this paper is foreign. Next, the design and implementation of the application layer, the ‘local cloud’ two-level storage mode, and the implementation of the cloud platform will be

```plaintext
4.

Overall Design of Medical Image Storage Architecture and Retrieval Application

In the case of large amount of medical image data, the data on the PACS system in the region is designed in a unified way, but it cannot meet the requirements of medical image data sharing. Through the analysis of the second chapter, the characteristics of HDFS can meet the requirements of medical image data storage and rapid processing, but it is not feasible to directly store DICOM3.0 standard image on HDFS, through the following discussion and analysis. HDFS and PACS systems have their own characteristics. HDFS can store Pb level massive data. When PACS is designed, it is limited to storing TB to hundreds of TB of data. HDFS also has the characteristics of low price, high reliability through data block copy, computing power in each storage node, and high throughput. Compared with PACS system, HDFS has its own defects. It is not a small file storage and cannot be applied in real time. Hadoop is a platform for storing and processing massive data under the current big data. As the core storage system of Hadoop, HDFS can satisfy all kinds of Internet big data processing companies such as Yahoo, Taobao and Baidu. Its big data storage characteristics can also satisfy the storage of medical image data. Combined with the actual application scenario, the paper builds the overall medical image storage architecture by combining the open-source his system, PACS system and distributed file system (introduced in Chapter 2). Combined with the application in the standard his system, it realizes the interface between the key his system, PACS system and distributed file system, and designs the “local cloud” two-level storage mode so that users can access his system directly access the local or cloud image files, and increase the cache to improve the overall access performance. The following will give a detailed introduction to each system and the overall architecture design. According to the long-term research and actual investigation, the use and test of two foreign open-source medical information management software, combined with the current situation of domestic medical image system, the design of Hadoop based medical image storage architecture is shown in Figure 3.

![Figure 3 Medical Image Storage Architecture](image_url)

According to the traditional mode used to manage various information data of hospitals in China, this architecture is composed of application layer, storage layer and platform layer, application layer is composed of clients of his system and PACS system, which is responsible for providing users with operation interface, information management, image viewing and other functions; storage layer is two-level storage mode of local end and cloud end, and local end is composed of his server and PACS server group which can be built on the local server, is responsible for storing and managing the hospital's structured information data and recent image data. The cloud is built by fast DFS large-scale distributed cluster, which is responsible for the permanent storage of long-term files. The data in the storage layer is uploaded by users and regularly uploaded from the local end to the cloud. The platform layer is built on the infrastructure by virtualization technology through the reasonable and effective use of server resources, it is convenient to provide cloud services. Most of the open-source software selected in this paper is foreign. Next, the design and implementation of the application layer, the ‘local cloud’ two-level storage mode, and the implementation of the cloud platform will be
```
introduced. GNU Health has almost undertaken all the contents of Tryton ERP. It consists of 12 modules. Each module can add a new function to the system. Through the complete installation of all the modules, they can support various business processes and related functions in the HIS system, as shown in Figure 4.

![Figure 4 Medical Image Retrieval Application Case](image)

5. Conclusion
In order to solve the characteristics of large amount of medical image data, combined with the current situation of big data research, through Hadoop big data cluster to solve the storage problem of massive image data. Combined with the PACS system of DICOM standard for medical institutions to store medical image data, a medical image storage system based on HDFS distributed file system is designed and developed, and the data model stored on HDFS is discussed, and the conversion mechanism of medical image data to system storage structure is realized.

1) First of all, this paper researches the medical image data and so on. Based on the research of the development of Hadoop technology, now there are many successful examples of Hadoop big data distributed cluster. Combined with the current characteristics of Hadoop use. And for the characteristics of HDFS trial, combined with the current status of medical image storage PACS system, in-depth research is carried out to lay the background for research experiments.

2) Through the basic knowledge, architecture platform and ecosystem of Hadoop cluster, what parts of Hadoop cluster are there, and the function that can coordinate the use of its projects in the future, so as to better use the big data cluster function. HDFS distributed file system with massive data storage capacity, its own system architecture characteristics and how to store data. And how to read and write the data in HDFS, and analyze the storage characteristics of HDFS, and then use HDFS to store medical image data to solve the problem of data structure model of its own medical image data in HDFS storage system.

3) Through the cognition of the current medical image storage system, we can understand the structure of the international standard DICOM file, the file data set stored in the file structure, how to store a large amount of data, and the format structure of the data set. It lays a good foundation for the following research on the storage of medical image files.

Acknowledgement
2019 Jingmen Guiding Scientific Research Project “Research of Medical Image Encryption on Scrambling and composite segmentation chaotic mapping based on dynamic update”, Number: 2019YDKY078.

References
[1] Norcen R, Podesser M, Pommer A, Schmidt H-P and Uhl A, Confidential storage and transmission
of medical image data Computers in Biology and Medicine, 2003, 33, 277-292.

[2] Acharya U R, Acharya D, Bhat P S and Niranjan U, Compact storage of medical images with patient information IEEE Transactions on Information Technology in Biomedicine, 2001, 5, 320-323.

[3] Acharya R, Niranjan U, Iyengar S S, Kannathal N and Min L C, Simultaneous storage of patient information with medical images in the frequency domain Computer methods and programs in biomedicine, 2004, 76, 13-19.

[4] Hani A F M, Paputungan I V, Hassan M F, Asirvadam V S and Daharus M, Development of private cloud storage for medical image research data. In: 2014 International Conference on Computer and Information Sciences (ICCOINS): IEEE, 2014, pp. 1-6.

[5] Vavilapalli V K, Murthy A C, Douglas C, Agarwal S, Konar M, Evans R, Graves T, Lowe J, Shah H and Seth S, Apache hadoop yarn: Yet another resource negotiator. In: Proceedings of the 4th annual Symposium on Cloud Computing: ACM, 2013, p.5

[6] O’Driscoll A, Daugelaite J and Sleator R D, ‘Big data’, Hadoop and cloud computing in genomics Journal of biomedical informatics, 2013, 46, 774-781

[7] Yang X Y, Liu Z and Fu Y, MapReduce as a programming model for association rules algorithm on Hadoop. In: The 3rd International Conference on Information Sciences and Interaction Sciences: IEEE, 2010, pp. 99-102