Research and Practice on Distributed Decentralized Medical Document Storage System

LIANG Bing-jin[1,a], TU Rui-wei[1], LIN Yan-jun[2,b*]

[1]Software School, Guangdong Food and Drug Vocational College, Guangzhou 510520, China,
[2]Department of Neurology, The Seventh Affiliated Hospital, Sun Yat-sen University, Shenzhen 518107, China

[a]liangbj@gdyzy.edu.cn, *[b]linyanjun@sysush.com

Abstract. In response to the security and stability issues of medical documents in large hospitals, this article designs a distributed decentralized medical document storage system based on IPFS to solve the problems of network congestion, vulnerability to virus attacks, poor security and confidentiality, and easy data changes caused by centralized storage of medical documents. The development and application of the system effectively enhances the patient's medical experience, and has important application scenarios and economic value in clinical diagnosis, medical research, and other fields.

Keywords: Medical documentation; IPFS; Distributed Storage Architecture

1 Introduction

With the continuous improvement of the level of medical construction in China, it has brought about the continuous growth of medical data such as electronic medical records, medical imaging, and medical video materials. The explosive growth of medical data is also driving the continuous improvement of medical informatization level[1].

The storage methods of electronic medical records are very diverse. Domestic hospitals all adopt centralized storage mode, generally using three storage architectures: NAS, DAS, and SAN[2]. But with the passage of time, the amount of data increases, and the scalability performance is poor. The cost of hardware equipment is also very high, and it is difficult to achieve sharing within hospitals or even regions, resulting in poor flexibility. The requirement for archiving and storing image information is to be able to meet the requirements of fast storage and convenient retrieval of a large amount of data. Therefore, choosing a suitable storage solution is crucial.

At present, most academic research on cloud storage of medical record files[3] uses HDFS as the underlying distributed file system. It supports offline batch data processing, supports high throughput, and is stored in data blocks. It can save multiple copies, which is easy to achieve load balancing. However, HDFS is suitable for processing small file data, and the smallest block is 64M, which is not suitable for small medical record files, Improving HDFS[4] for small files also increases engineering costs.

A. Rauf et al. (eds.), *Proceedings of the 3rd International Conference on Management Science and Software Engineering (ICMSSE 2023)*, Atlantis Highlights in Engineering 20,
https://doi.org/10.2991/978-94-6463-262-0_8

InterPlanetary File System[5] is a peer-to-peer distributed file system used to connect all computing devices using the same file system. IPFS is a peer-to-peer system where IPFS nodes store IPFS objects locally, and nodes connect to each other and transfer objects. IPFS provides a high throughput, content addressable block storage model. IPFS does not have a separate failed node and has advantages such as higher security, manageability, and scalability.

Many scholars have conducted research on the storage of medical electronic medical record systems and have good application prospects in the medical field. This project proposes a distributed decentralized medical electronic medical record storage architecture using IPFS as the underlying protocol.

2 Technology Introduction

2.1 P2P Network

P2P network[6], also known as peer-to-peer network, belongs to the coverage layer network and is a relative approach to the C/S mode. The C/S mode adopts a central node dedicated to distributing data, where clients can only obtain the required data information from the server and exchange data with other clients only through the server. This mode has good control over consistency issues, but when a single server encounters problems, system services will crash, and the requirements for network bandwidth and server CPU are relatively high. In addition, if a client exchanges data with other clients, it can only be done through the central server.

P2P has addressed these issues. In this network, all nodes can provide services to other nodes. At the same time, services can also be obtained from other nodes, and massive resources are utilized, no longer relying on the central end, and the central end is replaced. The blockchain[7] utilizes P2P technology, and when users want to synchronize a node, they need to first open the client. The client code has built-in information about several stable nodes, and when the client starts, relevant information will be obtained from these nodes. After the node starts synchronizing data, it will synchronize the information of other nodes from the connected nodes. As long as the proportion of failed nodes does not exceed a certain value, the impact on the entire system is not significant.

Decentralization of P2P technology. Its network does not have a central end in the C/S mode, and data exchange can be directly carried out between nodes. All resources and services on the network will be shared, and a large amount of network terminal resources will be effectively utilized.

P2P networks have strong scalability. As the number of users increases, the number of hosts connected to P2P networks increases, and the robustness of network strength also improves. The more nodes on the network, the more services it can provide. The scalability of the entire P2P network is theoretically unlimited.

The entire P2P network has a peer-to-peer relationship between each node. If some nodes commit crimes and maliciously damage resources, or if some nodes are attacked, in most cases, the impact on the entire network can be ignored. Unless the number of nodes responsible for wrongdoing accounts for the majority, it can have an impact on the entire network. As long as a portion of nodes can operate normally, the entire system can continue to maintain.

2.2 IPFS

InterPlanetary File System(IPFS) is a peer-to-peer distributed file system designed to connect all computing devices using the same file system. In some cases, IPFS is very similar to the Web.

IPFS provides a high throughput content addressable block storage model with content addressed hyperlinks. This forms the generalized Merkle DAG data structure, which can be built into a file version system, blockchain, or even a permanent web. IPFS uses distributed hash tables, incentivized block trading models, and self certified namespaces. IPFS does not have a separate failed node, and there is no need for a complete trust foundation between nodes.

IPFS is a peer-to-peer system where no nodes have privileges. IPFS nodes store IPFS objects locally. Nodes are connected to each other and transmit objects.

The advantages of IPFS lie in its strong technological accumulation, exquisite architecture design, and strong developer ecosystem. IPFS technology is divided into seven layers of protocol stacks from top to bottom: identity layer, network layer, routing layer, switching layer, object layer, file layer, naming layer, etc. The IPFS protocol stack is shown in the following figure. The underlying technology of IPFS consists of Distributed Hash Table (DHT), Block Exchange Protocol (BitTorrent), Version Control (Git), Self Verified File System (SFS), and others.

IPFS provides a high throughput, content addressable block storage model. IPFS does not have a separate failed node and has advantages such as higher security, manageability, and scalability. IPFS and blockchain can be perfectly matched, storing blockchain data in IPFS and storing IPFS links in blockchain practices to solve blockchain storage bottlenecks.

Many scholars have conducted research on the storage of medical electronic medical record systems and have good application prospects in the medical field. This project proposes a distributed decentralized medical document storage architecture that effectively prevents medical image data from being mistakenly deleted or maliciously attacked by humans or viruses, making the data more secure.

3 System Architecture

This project designs a distributed decentralized medical document storage system, taking the medical image storage architecture as an example, to build an IPFS private network within the hospital. In the radiology department, the medical imaging primary database server (which can have multiple servers), the information department secondary database server (which can deploy system servers such as WINDOW, LINUX, UNIX, etc.), and the imaging department workstation are deployed as IPFS private network nodes. System data flow chart is shown in figure 1 below.

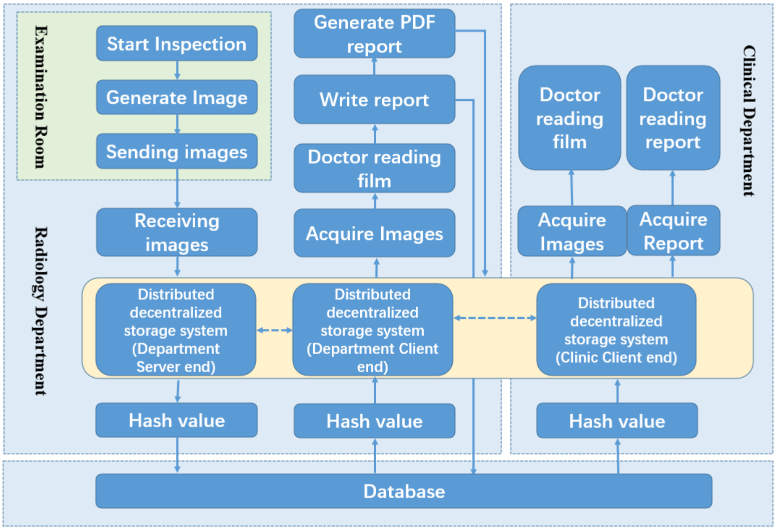

Fig. 1. System Data Flow Chart

3.1 **Central library server.** As the core (primary) server of the storage network, it is used to deploy an IPFS image storage server, which stores all medical document data in the hospital and can be accessed and obtained by clients from all departments of the hospital.

3.2 **Department server.** Deployed in radiology and other departments as an expansion node for IPFS departments, providing access to data for clients in the department.

3.3 **Radiology imaging workstation.** As an IPFS client, obtain medical images and other medical documents, mainly from department servers, supplemented by central libraries.

3.4 **Clinical workstation.** As an IPFS client, connect to the central library and obtain medical document data from the central library.

4 System Practice

The traditional medical image archiving process involves patients completing medical image examinations and generating medical images on the device. After receiving the images, the server stores them directly in the image library and stores them as image files. Unlike traditional medical image archiving, the server of this project receives images and directly stores them in the IPFS storage system. The archiving process is as follows.

(1) The device image processing workstation performs post-processing on medical images and transmits them to the DICOM server of PACS through DICOM services.

(2) After receiving the image, the DICOM server pushes it to the IPFS network, as shown in Fig.2 below.

```
(000001):CallingAETitle(startscu,127.0.0.1)/CalledAETitle(startpacs)
(000001):Receive C-STORE Request
(000001):Received image data :CT00005809-Ms Li
(000001):Save image to cache directory :image\CT\20141106\CT00005809_1.2.840.113820.491.2
(000001):Save image:0010_000153_1.3.12.2.1107.5.1.4.54693.30000006101906583670З
(000001):Receive C-STORE Request
(000001):Save image:0010_000154_1.3.12.2.1107.5,1.4.54693.30000006101906583670З
(000001):Receive C-STORE Request
(000001):Receive:0010 000155 1.3.122.1107.5,1.4.54693,30000006101906583670З
```

Fig. 2. DICOM Server Received Image File

(3) After receiving the DICOM file, the IPFS network returns the HASH encoding value of the DICOM file, as shown in Fig.3 below.

```
 58.00 MiB / 273.02 MiB [========>----------------------------]  21.24% 00m02s
added QmcuDFCBGzVAN19tYKYpzn9W3LWLngANKvkDQ9tCFHViTT XA/0001_110000_1.2.392.2000
36.9116.3.1.21002062.7.0.314844449125891
```

Fig. 3. HASH value returned after uploading the image medical record file

(4) After receiving the HASH value of the image, the PACS system stores the value on the database server; And notify the IPFS node of the secondary database to synchronize the image to the local node.

The process for obtaining medical images of patients is as follows.

(1) The PACS imaging workstation in the imaging department retrieves the patient's information and obtains the HASH value of the corresponding medical image for this examination, as shown in Fig.4 below.

```
4001/p2p/QmfBNkmUkFsZzqgE6DkrmWZZMuV6UVAcxhAzon6i4k19PF
4001/p2p/QmS8PoSUt5BTucK4oMMPLVBma9kAG4mfEqeN4yX257yZnU
4001/p2p/QmWXrZ4wuEYD9kUZcC89Ga7h3mWJmeb4wTkfkuCVNV89oS
4001/p2p/QmZj9HeD7rhmbTmxRBG2vqfLXXre5bKmA2AunJVXaqbHFT
4001/p2p/QmUtpxZJCpwSCDSss31TBTyJ89tsPx4L5kma22yMTgKyG4
4001/p2p/QmPXLy2ofumn8zzVR2Y3sszLMjmwisGsygN3kjc5v11Jwa
4001/p2p/QmRd3n5C9b4NZmn6xq18nXbc2Epj5WdCuZ7c5U6TiAszuE
4001/p2p/QmPwB4YJ81ZQwwgSghJeUJECjNg5Wrh2WcvrLtKkjn2Axj
4001/p2p/Qma3QRfjYMPKiiAp6Y3btEZhPtrwdHrB37Lt3njTyrAzqC
4001/p2p/QmQu7qfbMDXyrrzziNUanMBHyPJLkfu8heRpqL66yrHBue
4001/p2p/QmURhnR2dPmfcTjdbZcKUawuHxUy7nQ16SnHdb21xVA8UM
```

Fig. 4. List of HASH values in medical record documents

(2) The imaging workstation can directly obtain and display images through the local IPFS interface, as shown in the following figure, as shown in Fig.5 below.

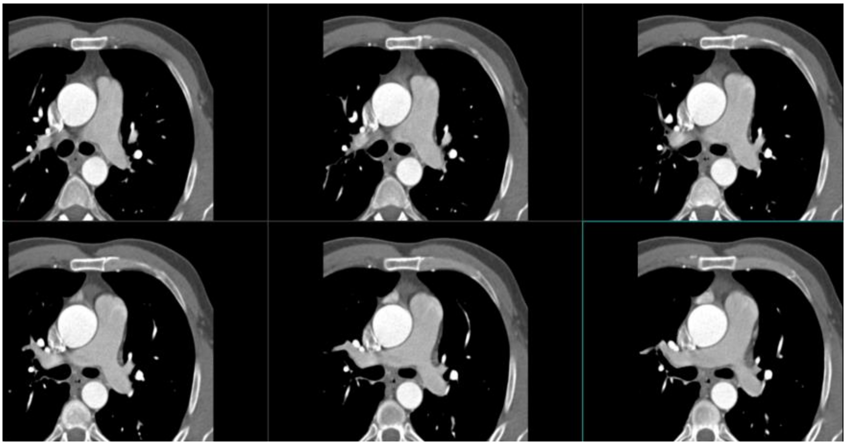

Fig. 5. List of HASH values in medical record documents

(3) The diagnostic doctor makes a diagnosis based on the patient's image and completes the diagnostic report. Diagnostic report file as shown in Fig.6 below.

CT Inspection Report

Name : TestPatient	**Sex**: Female	**Age** : 20	**PatientID** : CT00005811
Source : Obstetrics Clinic	**InPatientNo**:	**BedNo**:	**OutPatientNo**:

Checkpoint :

Clinical diagnosis :

Descripton:

The size and morphology of the liver are normal, the proportion of each lobe is consistent, the liver fissure is not wide, the capsule is smooth, and there is no obvious abnormal density in the liver parenchyma. The blood vessels in the liver are running normally, and the intrahepatic and extrahepatic bile ducts are not dilated. The spleen is not large, the gallbladder is not large, and the size, morphology, and density of the pancreas are normal. Both kidneys are symmetrical, and the size and morphology are normal. No focal density abnormalities are found. There are no bile lymph nodes in the retroperitoneum, and no abnormal enhancement is found after enhanced scanning, Bilateral ureteral running area not evident.

Impression:

1. No obvious abnormalities were found in the liver, gallbladder, spleen, pancreas, and both kidneys. 2. No obvious abnormalities found in the pelvis.

Fig. 6. Diagnostic Report File

(4) The diagnostic doctor makes a diagnosis based on the patient's image and completes the diagnostic report.

Clinical medical imaging workstation, which retrieves images and diagnostic reports through the clinical WebPacs workstation of the PACS system. WebPacs directly calls the image library IPFS node image, reducing the time for image copying back and forth and bandwidth consumption.

5 Conclusion

The main objective of this project is to construct a distributed decentralized medical electronic medical record storage system.

(1) The main objective of this project is to construct a distributed decentralized medical electronic medical record storage system.

(2) IPFS is stored in the node in block mode, which is not vulnerable to virus intrusion. It can effectively prevent medical electronic medical records from being tampered by humans or viruses, and is more resistant to computer virus intrusion.

(3) After storing medical electronic medical records in the IPFS network, the hash value of the electronic medical record file is returned. To obtain the electronic medical record file from a node in the IPFS network in the future, it must be obtained through the hash value. Effectively preventing malicious attacks, making images more secure.

With the increasing amount of data in electronic medical records and the multi hospital architecture, the pressure of storage and retrieval is increasing. Under the current centralized storage architecture, medical documents have low privacy and are prone to risks such as virus damage and copying. Moreover, with centralized storage, there are issues such as lag when multiple patients' images are viewed simultaneously, which seriously affects the speed of film viewing and leads to poor medical experience for patients. Therefore, this project aims to study the distributed decentralized storage architecture and solve the problems of traditional storage architecture. It aims to decentralize patient medical document storage, block encrypted storage to prevent copying and tampering, access to nearby areas, reduce disk IO and network bandwidth of the central server, store and access medical record files, optimize patient medical experience, and have good application prospects.

Acknowledgments

This research is supported by Medical Scientific Research Foundation of Guangdong Province (No: B2023452). and the Fundamental Research Funds of Guangdong Food and Drug Vocational College (No: 2021ZR07). The corresponding author is Lin Yanjun.

References

1. Sultanov M M , Gorban Y A , Smirnov A A , et al. Development of a centralized system for data storage and processing on operation modes and reliability indicators of power equipment[C]// 2021 3rd International Youth Conference on Radio Electronics, Electrical and Power Engineering (REEPE). 2021.

2. Chaudhari S , Swain G . Towards Lightweight Provable Data Possession for Cloud Storage Using Indistinguishability Obfuscation[J]. IEEE Access, 2022(10-):10.
3. Longwill S , Moore M , Flitcroft D I , et al. Using electronic medical record data to establish and monitor the distribution of refractive errors[J]. Journal of optometry, 2022, 15 Suppl 1:S32-S42.
4. Shvachko K , Kuang H , Radia S , et al. The Hadoop Distributed File System[C]// IEEE Symposium on Mass Storage Systems & Technologies. IEEE, 2010.
5. Li L , Yue Z , Wu G . Electronic Medical Record Sharing System Based on Hyperledger Fabric and InterPlanetary File System[C]// ICCDA 2021: 2021 The 5th International Conference on Compute and Data Analysis. 2021.
6. Satoshi M , Kazutoshi F , Hideki S . Mill: Scalable Area Management for P2P Network based on Geographical Location[J]. Proc Euromedia, 2022.
7. Kosba A , Miller A , Shi E , et al. Hawk: The Blockchain Model of Cryptography and Privacy-Preserving Smart Contracts[C]// 2016 IEEE Symposium on Security and Privacy (SP). IEEE, 2016.