Research on Medical Big Data of Health Management Platform Based on Hadoop

Xuan Zheng， Xiaopan Ding
College of Medical Information Engineering
GanNan Medical University
GanZhou, JiangXi
Xuanzheng@gmu.edu.cn; 820729509@qq.com

***Abstract*—The digitization process in the medical field continues to advance in depth and gradually enters the stage of smart medical treatment. Starting from the construction of the medical big data of health platform, this paper aims at the current problems of massive medical heterogeneous data resources and sharing, the difficulty of effective fusion of multi-source heterogeneous information in intelligent health services, and the real-time early warning of diseases, combined with the necessity of Hadoop platform and clinical medical data sharing, the HDFS distributed file system is used to manage metadata to improve the efficiency of integrated management of massive data; the data sharing mode of HBase is adopted to improve the degree of data sharing, with the HBase engine, Hive engine, SQLServer engine, etc., to meet the intelligence of different business data between heterogeneous, establish a data center with storage and management of patient health records and disease early warning as the core, and propose a big data management platform based on Hadoop to promote the scientific, professional, digital and refined operation and management of hospitals.**

Keywords- Hadoop; Big Data of Health; Distributed data sharing

I. INTRODUCTION

In 2017, the State Council issued the Notice on the Development Plan for a New Generation of Artificial Intelligence to develop artificial intelligence technology and explore the construction of smart hospitals. In 2018, it has successively issued the "Opinions of the General Office of the State Council on Promoting the Development of "Internet + Medical Health"" and the "Notice of the General Office of the National Health Commission on Further Improving the Appointment Diagnosis and Treatment System and Strengthening the Construction of Smart Hospitals" in 2020, which clearly focus on promoting "improving the service experience of the masses", "promoting the reform of medical insurance payment methods", "improving the universal medical insurance system, improving the management and service of medical settlement in different places, and basically achieving full coverage of ordinary outpatient expenses across provinces and direct settlement and coordination areas".

At present, the following problems are common in the construction of medical informatization in China: First, the scale of medical heterogeneous data resources is growing exponentially at the rate of tb and pb. Second, healthcare data involves a large amount of structured and unstructured data, and traditional data management systems are very difficult to effectively store and maintain data in a single hard drive, only processing megabytes and gigabytes of data, and when the data increases, performance decreases and data scaling is not supported. Third, the distribution and utilization of clinical business data is unreasonable. Clinical business data is not shared and centralized, resulting in clinical business data with huge medical value cannot be reasonably and effectively utilized.

However, one solution that can be solved is to use the Apache Hadoop framework, which provides a distributed file system and a scalable computing framework, and the advent of new technologies such as Hadoop has made it possible to store and process large amounts of data by dividing the computing process on many host servers (non-essential high-performance computers). Therefore, a big data processing architecture based on the Hadoop cloud platform can solve these problems.

With the continuous development and application of emerging technologies such as big data, cloud computing, and the Internet of Things, hospital informatization construction has the ability of active perception and intelligent regulation, which has a profound impact on the health service model. Relying on the medical and health big data platform, give full play to the advantages of medical and health big data collection, and build convenient and beneficial business application systems such as medical services, cost monitoring, drug management, and three-doctor linkage.

In order to achieve the above goals, this paper proposes a medical and health big data management platform based on Hadoop, which can integrate the patient diagnosis and treatment information resources of different medical institutions and grass-roots public health service institutions in the region, and establish a data center with the storage and management of patient health records and diagnosis and treatment information as the core, so as to realize the health information resource sharing and business collaboration mechanism between medical institutions.

II. MEDICAL BIG DATA OF HEALTH MANAGEMENT PLATFORM BASED ON HADOOP

Based on the Hadoop platform, the health and medical big data platform provides storage, integration, management, exchange, query, analysis, mining and other applications to support basic platform services and data processing, and adopts safe, convenient and efficient cloud computing and cloud storage solutions to provide

2473-3636/22/$31.00 ©2022 IEEE
DOI 10.1109/DCABES57229.2022.00062

intelligent and personalized "whole-person and whole-process" health care services. Provide data storage space and computing resources for regional health bureaus, medical and health institutions at all levels, and public health institutions. In order to meet the needs of different services for data reading speed, a hot and cold data storage system has been established to dynamically manage data according to the activity of data. The Hadoop analytics cluster completes parallel analysis of unstructured data and reports the results of the analysis. The parallel computing module mainly completes the storage, indexing, computing analysis and Hadoop computing management operations of unstructured data, and realizes the rapid parallel computing of unstructured medical and health big data.

Hadoop is a distributed computing framework that uses low-cost hardware devices as a basis for processing massive amounts of data. The distributed file system HDFS is used to manage metadata and build and store massive data; the data sharing mode of HBase is used to improve data sharing; Sqoop and Flume provide RDBMS data import functions for HBase, and it is very convenient to migrate traditional database data to HBase. MapReduce implements cluster distributed computing, using a distributed Spark cluster, accessing the machines in the cluster through the YARN manager, and the size of the cluster can be flexibly increased or decreased according to the size of the data processing volume. Therefore, the system implements the construction of a medical and health big data management platform based on the Hadoop platform, and its overall architecture is shown in Figure 1.

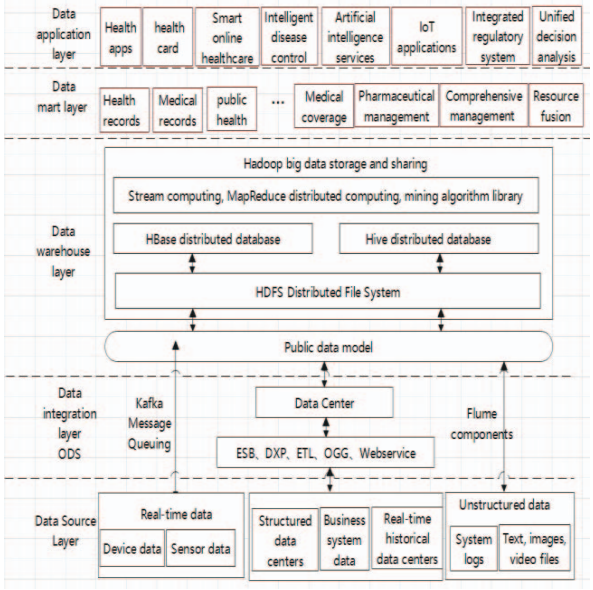

Figure 1. Hadoop-based medical data platform.

The platform is composed of a data source layer, a data integration layer, a data warehouse layer, a data mart layer, and a data application layer, which can provide data access, data integration, data sharing, data mining and analysis and other functions for related services.

A. HDFS-based Medical metadata management solutions

In the medical big data management platform built in this paper, a metadata management scheme based on the HDFS distributed file system is adopted, as shown in Figure 2. The second name node is an important part of the HDFS architecture and has two functions. First of all, the merge operation of EditLog and FsImage can be completed, which can reduce the file size of EditLog and shorten the restart time of the name node; secondly, it can be used as a "checkpoint" of the name node to save the metadata information in the name node.

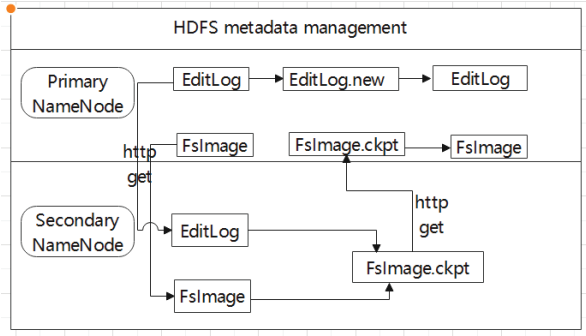

Figure 2. Medical metadata management solutions

The management scheme adopts the dynamic integration method of image file + operation log file, and the problem caused by the gradual expansion of EditLog can be effectively solved through the second name node. The specific process is as follows: (1) every once in a while, the second name node will communicate with the name node, request to stop using the EditLog file, and temporarily add the newly arrived write operation to a new file EditLog.new; (2) the second name node obtains the FsImage and EditLog files from the name node through Http get and downloads them to the corresponding directory locally ;(3) The second name node loads the downloaded FsImage into memory, and then performs the update operations in the EditLog file one by one to keep the FsImage in the memory up to date, the process is to merge editLog and FsImage file, and then send the new FsImage file to the name node by post ;(4) NameNode replaces the old FsImage file with the new FsImage received from the second name node, and replaces the EditLog file with EditLog.new, making the EditLog smaller.

Through the above method, the second name node is equivalent to setting a "checkpoint" for the name node, periodically backing up the metadata information in the name node, and being able to rely on the Second NameNode file to achieve data recovery in the event of a system failure. Through the metadata management system, the attributes of the data can be analyzed, so that the number is active and the number is evidenced, and the standardization, completeness and accuracy of the data are guaranteed.

B. Distributed Data Sharing Based on HBase

Medical big data is the foundation of smart medical care and the core point that connects various intelligent technologies, new products and users. Taking data demand as the dimension, using distributed storage to centralize the hospital's data, storing it in the database through desensitization, cleaning, structuring, normalization and quality control, etc., and establishing a platform for data management, analysis and application on this basis, which

is the basic idea of building a medical big data platform. The medical intelligence platform based on big data is conducive to collecting and processing data, obtaining disease-related information in a timely manner, and realizing automatic diagnosis, early warning and prevention measures for diseases through technologies such as sensors and artificial intelligence, so as to play a role in providing personalized medical services and popularizing high-quality medical resources. However, in view of the uneven development of the current level of hospital informatization in China, the electronic medical records, image archiving, communication systems and clinical information systems have not yet been popularized, and there are still defects in the security management of relevant information and personal privacy, although most hospitals have established information systems, but due to lagging standardization, data sharing is seriously affected.

The data warehouse layer based on the open source Hadoop framework is responsible for sharing and exchanging heterogeneous data of different business systems, and the data sharing mode adopted is mainly used by HBase as a bridge for data sharing, establishing a virtual mapping relationship between global and nodes, and using a variety of integration technologies to achieve centralized control and high sharing of data, as shown in Figure 3.

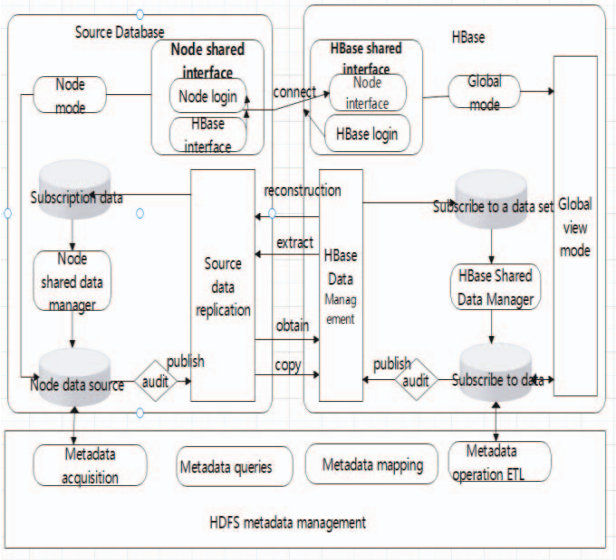

Figure 3. HBase-based data sharing model

C. *Real-time analysis of healthcare and disease warning*

Health early warning research pays special attention to diseases with a large number of patients, high mortality rate, and a major impact on the life and economy of the people, so it chooses to take cardiovascular and cerebrovascular diseases, diabetes, etc. as examples to carry out high-precision rapid warning model research on diseases. Relying on the medical big data management platform to dynamically manage patient data, intelligent analysis, tracking and control of the treatment process, and fully guarantee medical safety.

Individualized features of disease development can be found from vital sign monitoring data. First of all, for each disease, according to medical knowledge and expert experience, some key signs and indicators that are closely related to the development or diagnosis of each disease need to be identified, such as cardiovascular and cerebrovascular diseases are closely related to blood pressure indicators, but there is no obvious correlation with the number of intestinal colonies. Secondly, the monitoring data of multi-indication indicators has the characteristics of multi-source and heterogeneous (different formats such as values, images, text, etc.), and multi-source data need to be processed and analyzed as necessary. Finally, because the patient's physique not only has personalized characteristics, but also changes with the patient's age, the season and the change of living environment, it has a certain dynamic. At the same time, through the mining of a large number of patient monitoring data, the general law of disease development and change can be found, and the combination of the two can form personalized disease prediction.

In order to provide patients with effective and economical health information services, it is necessary to assess the risk of onset according to the patient's medical history, health records, lifestyle, etc., and provide online early warning services of different degrees and categories for users with different risks. Therefore, the knowledge representation and knowledge base design of health and medical information are established, and a big data integration method of health service management and decision-making is established to provide an effective technical implementation path for health service management.

III. PLATFORM TECHNOLOGY ARCHITECTURE

The platform technology architecture consists of four levels: Infrastructure as a Service (Iaas), Data Resource as a Service (DaaS), Platform as a Service (Paas), and Software as a Service (Saas).

The IaaS layer is located at the bottom of the four layers of services of the cloud platform. It relies on the cloud environment, storage resources and network resources to provide the most basic computing resources, including CPU, memory, servers, storage, etc., and abstracts these resources into externally available services. The provision of services is a key link, and the quality of services directly affects the user's use efficiency and IaaS system operation and maintenance costs.

The DaaS layer is located above the IaaS layer. It provides scalable, highly available, and multi-tenant database services for platform application buildings. Integrate all kinds of data from medical institutions, public health institutions and other related systems at all levels to achieve centralized data management and ensure service levels. It mainly includes three parts: data collection and storage, data processing and data service provision, and the specific technologies include ETL, distributed storage, and distributed computing.

The PaaS layer is located on top of the DaaS layer of the cloud platform's four-tier services. It provides an Internet-based application development environment for end users, including application programming interfaces and operating platforms. The PaaS layer supports a variety of applications required throughout the lifecycle, from

creation to operation of software and hardware resources and tools.

The SaaS layer is located on the top of the four layers of services of the cloud platform. Users can use the application software deployed on the cloud platform through various portals such as PC terminals, mobile terminals, and official accounts. Service providers provide consumer or industry applications directly to end users and various enterprise users.

IV. THE APPLICATION EFFECTIVENESS OF THE MEDICAL BIG DATA MANAGEMENT PLATFORM

Hospital health and medical big data is designed with the purpose of "health management and patient-centered", making full use of the collected data for effective analysis, and in the early stage of system application, it can integrate the scattered and disorderly medical and health information data, and automatically analyze according to machine learning and artificial intelligence mode to form a transparent and intuitive data analysis report. After actual use and research, the use of the questionnaire survey evaluation system of the hospital doctor-patient app mobile phone platform, the objective investigation of medical staff and patient satisfaction, from the big data mining and its statistical analysis data tools before and after the use of comparison, the resident health management file filing rate increased from the original 56% to 91%, the patient's satisfaction with the hospital increased from the original 83% to 95%, and the early warning and early screening rate of related diseases increased from the original 50% to 75%. The use of big data technology can not only effectively reduce medical costs, but also integrate patient genetic information to guide personalized treatment, and use big data technology to analyze population health data can also prevent disease outbreaks.

A convenient and beneficial service system has been established. Relying on the health and medical big data platform, it gives full play to the extensive connection characteristics of "Internet + medical health". Encourage health service personnel to carry out online medical services such as online consultation, online follow-up, appointment examination, viewing results, online follow-up, prescription, and drug delivery. In public health emergencies, relying on artificial intelligence assistance systems, it is possible to quickly help key patients such as fever at the grass-roots level or designated hospitals to check for fever and carry out orderly and graded treatment. At the same time, as an important supplement to medical resources, the health and medical big data platform expands the supply of medical resources and forms a service system that benefits the people.

Improved dispatching and command capabilities. Through the health and medical big data platform, it can not only realize the information collection, dynamic monitoring and command and dispatch of public health events, but also realize the information transmission and information resource sharing of various departments. The platform can scientifically dispatch emergency health resources, display the distribution, demand and availability of urban health resources from multiple angles, facilitate the query, tracking, analysis and traceability of resource utilization, and maximize the value of resource use. In addition, the platform can also provide decision-making basis and command tools, providing situation research, judgment information and analysis methods for health emergency department operators and experts.

Surveillance and early warning of infectious diseases have been carried out. Based on health records and electronic medical records, relying on the infectious disease case data and symptom monitoring data collected by the platform, the comprehensive use of big data and artificial intelligence technology is used to build an epidemic decision-making command model. Then, according to the needs of disease control business monitoring, research, judgment and disposal, different types of data are automatically generated to analyze and predict the trend of the epidemic.

ACKNOWLEDGMENT

Project funds: Department of Education Science and Technology Project in JiangXi Province - Research on Medical Big Data of Health Management Platform Based on Hadoop (GJJ190815).

REFERENCES

[1] Shafqat, S., Kishwer, S., Rasool, R.U. et al. Big data analytics enhanced healthcare systems: a review. J Supercomput 2020,76:1754–1799.

[2] Choi, S., Chung, K. Knowledge process of health big data using MapReducebased associative mining. Pers Ubiquit Comput 2020(24):571–581.

[3] Satti F A , Ali T , Hussain J , et al. Ubiquitous Health Profile (UHPr): a big data curation platform for supporting health data interoperability[J]. Computing, 2020(2).

[4] Lv Z , Qiao L . Analysis of healthcare big data[J]. Future Generation Computer Systems, 2020, 109:103-110.

[5] Natalia Shakhovskaa,Solomia , et al. Big Data analysis in development of personalized medical system, Procedia Computer Science, 2019, 160:229-234

[6] Sadooghi I,Martin JH,Li T,et al.Understanding the performance and potential of cloud computing for scientific applications[J].IEEE Trans Cloud Comput,2015,5(99):1.

[7] Abbasi A A , Abbasi A , Shamshirband S , et al. Software-defined Cloud Computing: A Systematic Review on Latest Trends and Developments[J]. IEEE Access, 2019, 7(99):93294-93314..

[8] WU C Y, AHMED A, BEUTEL A, et a1.Recurrent recommender networks proceedings of the 10th ACM international conference on web search and data mining[C]. Cambridge, 2017: 495-503.