The Model Design of Medical Data Life Cycle Based on Big Data Platform

Ziqing Li a, Shenglei Pei b, Guiliang Feng c

School of physics and electronic information engineering Qinghai University for Nationalities Qinghai, Xining, China

a liziqing@qhmu.edu.cn, b 7597508@qq.com, c 371761995@qq.com

Abstract. With the gradual popularization of the mobile Internet and the Internet of Things, people’s lifestyles have undergone profound changes. In the medical field, the demand for smart medical remote care is gradually increasing. The continuous upgrading and updating of medical equipment and the precision operation of inspection equipment have made medical information systems contain more and more types of data. Based on data lifecycle management theory and big data processing technology such as Hadoop, this paper designs a corresponding data lifecycle management model for children's chronic kidney disease remote rehabilitation system. Its role is to manage the data reasonably, thereby increasing the availability of information systems and improving the efficiency of system storage devices.

Keywords: Data life cycle management, Mobile medical system, Hadoop.

1. Introduction

With the rapid development of the Internet, more and more traditional industries have been injected with new scientific and technological vitality. In the medical industry, mobile medical and intelligent medical are gradually moving towards the stage of practical application. However, the explosive growth of information brought by the rapid development of the Internet challenges the traditional data storage management. In order to improve this situation, this paper uses the relevant theory of information lifecycle management and big data technology to upgrade the data management method of the platform, and establishes an information lifecycle management model that meets the needs of the target system.

Information lifecycle refers to the process of information and its metadata from generation to initial storage, classified storage, use, final archiving or deletion. According to the value of data, information lifecycle management is to classify and manage the data, realize the hierarchical storage of data, and finally realize the hierarchical automatic transformation of data. At present, there are many researches and practices about information lifecycle. After entering the 21st century, many companies have carried out research in this field. The concept of information lifecycle management (ilcm) is proposed by legato company, and the concept of data lifecycle management (DLM) is proposed by StorageTek company. The joining of these enterprises also improves the research speed of information lifecycle management.

In this case, the mobile medical platform involved in this case is a rehabilitation platform for children with chronic kidney disease. The original intention of the platform is to help children and medical personnel to get in touch with each other remotely and assist in rehabilitation guidance and other related
work. However, with the gradual increase and improvement of the function, the original data storage scheme is too single, which leads to the decline of system performance. Based on the theory of information lifecycle management and the characteristics of the system, this paper designs an information lifecycle management model, which improves the efficiency of system storage space and system performance.

2. Related concepts

2.1. Information Lifecycle Management
There is no widely accepted process of information lifecycle management, but it is generally believed that information lifecycle includes the following parts: information collection, information value evaluation, information storage, information archiving / deletion.

Information Collection refers to the information that enters the system and needs further processing by the system. The source of information can be external input of the system, such as wearable devices, Internet, etc., or the operation log generated by the system itself. These data are characterized by large amount of information.

Information value evaluation refers to making corresponding rules to evaluate data according to some characteristics of data, distinguishing high value data from low value data, so as to prepare for the next step of information storage. At present, there are many methods for data value evaluation. Literature [2] provides a method based on data access frequency, which requires less parameters and is easy to implement in the actual system [2]. Reference [3] provides another comprehensive evaluation method for data value evaluation based on data content, business value and its own attributes. Reference [4] provides a data evaluation method based on data block size, which evaluates the data value according to the throughput of disk and the access amount of data block [4]. Reference [5] provides a method of value evaluation using machine learning [5].

Information storage refers to the storage method of information in the system. Due to the development of hardware technology, in addition to the traditional HDD Hard disk, there is also a faster SSD. Different hardware devices have great differences in performance and price. Therefore, in order to make more effective use of the characteristics of different storage devices, hierarchical storage technology can reduce the cost of the system and optimize the performance of the system.

When the use value of information tends to zero, the information should be filed or deleted to provide storage space for new information. The process of information lifecycle management is basically composed of the above processes. In the actual operation process, the lifecycle may be adjusted according to the specific situation. For example, in reference [6], the author applied information lifecycle management to the information system of smart city [6], only designed three stages of data acquisition, data processing and data retention. The most important thing is to use the theory flexibly to serve the system.

2.2. Remote rehabilitation support system for children with chronic kidney disease
Chronic kidney disease (CKD) in children refers to chronic persistent renal dysfunction caused by a variety of causes, and eventually develops to end stage renal disease (ESRD) [7]. This disease is a long-term, chronic disease. Data from the United States Renal Data System (USRDS) show that the incidence rate of adult renal failure is increasing year by year, and the treatment cost of this disease is high and the treatment effect is poor [8]. If the disease can be diagnosed and treated in the early stage, the recovery rate of patients can be increased. The disease is a kind of chronic kidney disease, the treatment cycle is very long, and the patient's diet needs to be monitored for a long time. If the patient is far away from the hospital, the back and forth rush will also bring unnecessary trouble to the patient's family and life. Therefore, remote medical rehabilitation can reduce the traffic cost of patients, strengthen the communication between doctors and patients, and solve the problem of chronic kidney disease treatment and monitoring.
2.3. *Hadoop related technologies*

Hadoop is an open source distributed computing platform of Apache foundation, which provides users with related tools of distributed computing. With the continuous progress of big data technology, Hadoop has gradually included many big data subprojects, including MapReduce, HDFS, HBase, etc. In the Hadoop ecosystem, HDFS is the most basic. It allows the system to store super large files on multiple computers in a distributed way. In addition, Hadoop provides MapReduce technology, which can solve the efficiency problem of large-scale data processing process. MapReduce technology simplifies the large-scale parallel calculation into map function and reduce function, and greatly reduces the network overhead of data transmission by dividing the data into segments for processing. Developers usually only focus on the implementation of map function and reduce function to realize the operation of super large data.

3. **Model analysis and design**

3.1. **System structure**

The remote rehabilitation support system for children with chronic kidney disease consists of four terminals and three subsystems, which are doctor app, patient app, nurse app, intelligent water cup, and background server, patient condition auxiliary diagnosis system and food identification system. The functions of each part are shown in Figure 1.

![Fig. 1 Telemedicine system of children with chronic kidney disease](image)

Patient app: considering that the patient is younger (4-6 years old, generally not more than 12 years old), the main user of the patient app is the patient's guardian.

The functions of the app on the patient side include viewing medical records, checking doctor's orders, and checking their own movement and health data. In addition, patients also need to send their own drinking water records and eating pictures to the background server every day as reference data for doctors' rehabilitation training. In addition, the patient app can also make video calls with the doctor app and the nurse app to directly communicate their condition and rehabilitation plan and their own rehabilitation situation.

Doctor app: Doctor app is used by doctors. The doctor can retrieve the patient's medical records from the background server, and write medical orders. Doctors can also communicate with nurses through instant messaging (text chat) to facilitate the deployment of rehabilitation details. In addition, the doctor can also make video calls and instant messaging (text chat) with the patient.

Nurse app: nurse app is used by nurses. Through the app, nurses can arrange appointment consultation between doctors and patients, fill in patient's medical records, manage the subordinate relationship between doctors and patients, and make appointment for telephone diagnosis and treatment.
Background server: the background server stores most of the data of the remote rehabilitation system for children with chronic kidney disease, including patient's personal information, visit record, patient's laboratory report and patient's medical record. It also stores the account information of doctors, patients and nurses. In addition, the backstage also manages and stores the information of other data analysis, including the related data of data analysis and simulation diagnosis by using the patient's data, as well as the information of the patient's diet record and food picture information.

Patient app: considering that the patient is younger (4-6 years old, generally not more than 12 years old), the main user of the patient app is the patient's guardian. The functions of the app on the patient side include viewing medical records, checking doctor's orders, and checking their own movement and health data. In addition, patients also need to send their own drinking water records and eating pictures to the background server every day as reference data for doctors' rehabilitation training. In addition, the patient app can also make video calls with the doctor app and the nurse app to directly communicate their condition and rehabilitation plan and their own rehabilitation situation.

Doctor app: Doctor app is used by doctors. The doctor can retrieve the patient's medical records from the background server, and write medical orders. Doctors can also communicate with nurses through instant messaging (text chat) to facilitate the deployment of rehabilitation details. In addition, doctors can also make video calls and instant messaging (text chat) with patients

Nurse app: nurse app is used by nurses. Through the app, nurses can arrange appointment consultation between doctors and patients, fill in patient's medical records, manage the subordinate relationship between doctors and patients, and make appointment for telephone diagnosis and treatment.

Background server: the background server stores most of the data of the remote rehabilitation system for children with chronic kidney disease, including patient's personal information, visit record, patient's laboratory report and patient's medical record. It also stores the account information of doctors, patients and nurses. In addition, the backstage also manages and stores the information of other data analysis, including the related data of data analysis and simulation diagnosis by using the patient's data, as well as the information of the patient's diet record and food picture information.

Food identification system: due to the need to monitor the patient's daily diet, the system obtains the patient's food information through the food picture information uploaded by the patient, and identifies it through the food identification system. The system is a new system added in phase II and is still under improvement.

Intelligent water cup: in order to monitor the amount of drinking water for children patients, a special water cup is provided for children patients, and the daily water consumption of patients is detected by sensing the water volume change of the water cup. And through Bluetooth communication, the data is uploaded to the patient app.

The patient's condition is monitored by the neural network, and the patient's condition is monitored by the neural network.

3.2. System data type
From Figure 1, we can analyze the data interaction in the system, and then know the type of data and the owner of the data.

On this basis, we design the corresponding data lifecycle model.

The data types of the system are shown in Table 1.

| data type          | Data owner          |
|--------------------|---------------------|
| Text data          | Doctor / nurse / patient |
| Voice data         | doctor              |
| image data         | patient             |
| Electronic medical record | patient           |
| network data       | system              |
In this system, there are five types of data: text data, voice data, image data, electronic medical record and network data. Text data refers to the text information in the system, mainly the user's account number, password and other information. Voice information is the voice information produced by doctors when they write voice medical orders. The image information is the food image uploaded by the patient. Electronic medical record (EMR) is a kind of information in special format, which contains the patient's medical records, laboratory test sheet data and other information. Because of the complexity of the composition, these information are often viewed by patients and doctors, so they are listed separately. Network data information refers to the system log generated during the operation of the system, as well as the video cache information generated by video calls between users.

3.2.1. Information lifecycle management model. The information life cycle model of remote rehabilitation support system for children with chronic kidney disease is shown in Figure 2.

![Fig. 2 Information lifecycle management model](image)

In Figure 2, we divide information lifecycle management into three parts: data acquisition, data value evaluation, data processing. Data processing includes data storage and deletion, and the specific operation is determined according to the results of data value evaluation. First of all, when the program runs, it enters the data acquisition stage. After obtaining the data, it evaluates the value of the data and processes the data separately according to the value of the data. The high-value data is stored on the faster SSD data, and the low-value data is stored on the slow hard disk, and the unnecessary data is directly deleted. Each part is described below.
3.2.2. Ata acquisition. In the data acquisition phase, we will give priority to the new data entering the system. However, if it is the first time to evaluate the value of data, all data should be scanned and prepared for value evaluation. In the process of scanning data, the system mainly obtains the data type, data owner, data size, data age and data access frequency.

Data types refer to the five types listed in Table 1. We distinguish the five data types according to their priority.

| data type            | Data owner     | Data priority |
|----------------------|----------------|---------------|
| Text data            | Doctor / nurse / patient | 3/3/3         |
| Voice data           | doctor         | 2             |
| Image data           | patient        | 2             |
| Electronic medical record | patient    | 3             |
| Network data         | system         | 1             |

In the whole system, we divide the data into three levels according to their importance. The importance of data increases from 1 to 3. As can be seen from table 2, we set the priority of text data to 3. This is because the text data is basically the user's personal information and account information. These are important information, so set it to 3. Secondly, considering that the voice data information is only the doctor's voice medical order information, this part of information is less important than the text data, so it is set to 2. In this system, there is a lot of image information, which comes from the food photos of the patient app. This part of information is used for food identification, and its importance is weaker than the text data. Although EMR information is patient information, it is also a file that doctors and nurses often need to view and edit. Therefore, this part of information should be stored on faster hardware equipment, and the importance is 3. For the network data of inter system communication, including the system's own cache and operation log, the importance of this part of data is not high for the system, so the importance is set to 1.

3.2.3. Data value assessment. At the end of data acquisition, the data value is evaluated. In the process of data value evaluation, we mainly consider data priority, data age, data size and data access frequency. The formula of data value evaluation is as follows:

\[ V = (k)V_S + (1-k)V_D \]  \hspace{1cm} (1)

\[ V_S = \frac{1}{2} \left( \frac{1}{\text{size}} + \frac{1}{\text{age}} \right) \]  \hspace{1cm} (2)

\[ w(t) = \frac{1}{T_i} \]  \hspace{1cm} (3)

\[ q(t) = \frac{\text{Usage of data in phase } t}{\text{Total data usage}} \]  \hspace{1cm} (4)

\[ V_D = \frac{\sum_{i=0}^{N}(w(t) \cdot q(t))}{N+1} \]  \hspace{1cm} (5)

In formula (1), \( V \) is the value of data, \( V_S \) is the static value of data, \( V_D \) is the dynamic value of data, in formula (3), \( T_i \) represents the period of data access, the current period is \( T_0 = 1 \), and the last period is \( T_1 = 1 \). Static value refers to the part that has nothing to do with the access frequency, which is mainly affected by the age of the data (in weeks) and the size of the data (in KB). Dynamic value refers to the part related to data access frequency. Since access frequency is the main influencing factor, and the importance of access frequency changes with time, we call the value of this part of data determined by access frequency dynamic value. \( W (t) \) is the weight of access frequency. The closer the access is to the current period, the higher the value weight is. \( K \) is the proportional adjustment parameter. In this model, we let \( k = 0.5 \). After formula (1), we can get the value \( V \) of the data. But in order to highlight the importance of each data, we deal with \( V \) and get the final data value through formula (2).

\[ V_{\text{final}} = \alpha V \]  \hspace{1cm} (6)
In formula (6), \( \alpha \) represents the importance level of data. Because the formulas (2) and (5) normalize \( V_s \) and \( V_d \), the range of \( V \) is \([0,2]\). In the previous article, we graded the importance of data according to 1-3. We assign a value to \( \alpha \) according to Table 3.

| \( \alpha \) | Data importance |
|---------------|-----------------|
| 1             | 1               |
| 1.5           | 2               |
| 2             | 3               |

From Table 3, the final value of the data can be determined. In this case, by adjusting the parameter \( \alpha \), the important data can be stored on the fast device in a planned way.

3.2.4. Data processing. After the data value evaluation, it is necessary to classify the data according to the different values. There are four main ways of traditional data processing: online storage, near line storage, offline storage and deletion. In this model, considering the future increase of users and system functions, as well as the ease of deployment of cloud server, we simplify the data processing after value evaluation into three categories: online storage, near line storage and offline storage.

In the process of data value evaluation, after the data value is calculated, the data will be layered according to the importance and other attributes. According to the data value, we give different processing methods, as shown in Table 4.

| Data importance | \( \alpha \) | Value range | processing method |
|-----------------|--------------|-------------|-------------------|
| 1               | 1            | \([0.2]\)   | offline storage   |
| 2               | 1.5          | \([0.3]\)   | Near line storage |
| 3               | 2            | \([0.4]\)   | Online storage    |

4. Code implementation of the model
Considering the amount of data in the future system, Hadoop and Java are used as experimental tools to realize the information lifecycle management model. The experimental environment is shown in Table 5.

| Experimental equipment | operating system | Hadoop version |
|------------------------|------------------|----------------|
| DELL OPTIPLEX380       | Ubuntu 16.04     | 2.8.0          |
| LENOVO IDEAPAD 700     | Windows10        | 2.8.0          |

We have built three Hadoop clusters of Ubuntu, one namenode and two datanodes. The version of Hadoop is Apache distribution 2.8.0. Windows computer is used for Hadoop client programming. Due to space constraints, the detailed configuration steps are not listed here. Using the file status API and disk trace information provided by Hadoop, we can get the file name, data type, file size, creation date, and historical access frequency. Then we use Hadoop for client programming, data value evaluation and data storage. The code is too long, here only lists the pseudo code of the key data value evaluation, which is convenient for readers to refer to.

The implementation code is as follows:

```java
class DataValuation{
  dataSize;// data size
  dataAge;// Data age
}
DataType;// data type
DataUsage[N];// The data usage array divides the data into N equal parts according to the age,
DataUsage // Record the usage of each copy
Vs=(1/DataSize+1/DataAge)/2;
While (0≤i≤N-1)
Do{
Sum +=(1/N-i+1)*(DataUsage[N])
i++
}
Vd=Sum/N
V=DataType *(0.5*Vs+0.5*Vd)

In the above code, the size and age of data can be obtained in map stage through API provided in Hadoop, and the data usage needs to be obtained by trace tool. The data type can be judged by the data source.

5. Conclusion
In this paper, according to the actual case, the information lifecycle management theory is applied to the mobile medical industry, and combined with the characteristics of the system with large amount of data and various types of data in the future, we choose to use big data technology for development. Finally, the information lifecycle management is implemented on Hadoop platform with Java language.

In this paper, the data size, name, access frequency and data type are comprehensively judged, and the multi factor data value evaluation method is used to realize the value evaluation of the data. Although the system has not been used in the actual environment. However, the model can still be used as a reference for the construction of big data in the actual medical system and a new idea for medical institutions to manage their own medical data.

Acknowledgements
This work was supported by Research project of school level teaching reform in Qinghai University for Nationalities in 2020 (2020-JYQN-006): Application Research of online learning system based on Android and Applied basic research project of Qinghai Province (2019-ZJ-7017): Decision making and early warning of ecological animal husbandry development based on multimodal collaborative learning.

References
[1] Yang Haifeng. Introduction of information lifecycle management concept by legato [J]. Communication world, 2003:24:49
[2] Ying Chen, "Information Valuation for Information Lifecycle Management," Second International Conference on Autonomic Computing (ICAC’05), Seattle, WA, 2005
[3] Gauri Shah, Kaladhar Voruganti, Piyush Shivam, et al. ACE:Classification for Lifecycle Management.RJ10372(A0602-044).Armonk,NY:IBM,2006
[4] X. Zhao, Z. Li and L. Zeng, "A Hierarchical Storage Strategy Based on Block-Level Data Valuation," 2008 Fourth International Conference on Networked Computing and Advanced Information Management, Gyeongju,2008, pp. 36-41.
[5] Büsch S, Nissen V, Wünscher A. Automatic classification of data-warehouse-data for information lifecycle management using machine learning techniques[J]. Information Systems Frontiers, 2016:1-15.
[6] A. Sinaeepourfard, J. Garcia, X. Masip-Bruin and E. Marin-Tordera, "A Novel Architecture for Efficient Fog to Cloud Data Management in Smart Cities," 2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS), Atlanta, GA, 2017, pp. 2622-2623.
[7] Combination method of flexible evidences in D-S evidence theory [M]. Beijing: National Defense
[8] US Renal Data System. 2010 Annual Data Report: atlas of chronic kidney disease and end-stage renal disease in the United States [OL]. [2011-06-13]. Bethesda. http://www.usrds.org/adr.htm. 2011-06-13.