ISSN: 2688-9323

Research on Scientific Data Management Methods for Scientific Research Project Implementation

Miao Yan

School of economics and management, Xidian University, Xi'an 710000, China.

ym_breve_girl@163.com

Abstract

For scientific and efficient management of scientific research projects generated huge amounts of heterogeneous data, this paper studies the scientific research project in the process of scientific data management method, data management method of the existing research both at home and abroad on the basis of the combination of existing information technology, the distributed storage of scientific data of HDFS and store the metadata information of mongo, combining to form a new scientific data management method, this method can reduce disk file scientific data access frequently called IO process, improve the rate of scientific data access and obtain, for scientific data management in the field of data management provide reference and reference.

Keywords

Scientific data management, Scientific data management methods, HDFS distributed storage, MongoDB.

1. Introduction

Scientific data is an important strategic resources, the natural science foundation (NSF) can be divided into four types, observation data, the experimental data and calculated data and record data [1], in the process of scientific research, because of the lack of scientific management, make scientific data, the phenomenon of the loss, in addition, in the process of scientific data management another tricky question is, what are the long-term preservation of scientific data need? How long is the shelf life? Is the key of the scientific data management method in the research problem, put forward the implementation of scientific research project of scientific data management method, which can effectively solve data loss in the history of scientific research, but also solved the problem of the long-term preservation of scientific data, so as to avoid repeated research in scientific research, can greatly save manpower, financial resources, promoted the process of the subsequent scientific research, therefore, it is necessary to explore the method of scientific data management research.

Objects of scientific data management including research data, metadata[3, 4], scientific data management is a scientific data in the scientific research project life cycle to the research data, metadata management of all activities [2], the lifecycle data based on scientific data management activities include the following aspects: data collection, data analysis and processing, data storage, data mining, and Shared [5], and realize the above management activities, need certain technology and method [6].In scientific research data management method, related to the existing research results at home and abroad, such as Haitao Li believes that combines computer technology and Marine science data business, realize the management of the scientific data for Marine field [7], Lili Zhang also believe that the flexible application of information technology to the boundary of scientific data management is an important factor to promote the development of scientific data management [8], Jingran Ma think by creating such as indexes and sorting, establishing contact between tables, optimizing scientific data management method [9] the query speed. Abroad on the study of scientific data management method, pay more attention to practice, focus on the information such as data type, data, or metadata format [10], Fischer M think data such as after the leave data acquisition device, usually need to go through several processing steps, can achieve the data published results [11], in the study of scientific research generated huge amounts of heterogeneous data, Stillerman J think such as metadata can help researchers to quickly understand and use scientific data [12], Huang X pointed out in the paper that mass data and metadata describing mass data are stored in HDFS and HBase, through which users can query mass data [13].

Review described above, management method was studied for the science data in the process of scientific research project at home and abroad has become a hot topic in the field of scientific data management, domestic research in scientific data management method, pay more attention to the theoretical study, put forward the application of modern information technology to different disciplines in the field of scientific data management, through the establishment of the index at the same time, provides a quick query, foreign to the scientific research data management method, from theory research to practice step by step, more focus on the practice of the scientific data management research, put forward using the hadoop and HBase to store huge amounts of heterogeneous scientific data, At the same time, metadata corresponding to scientific data is stored, and mass data is queried through metadata information, so as to access and obtain mass heterogeneous scientific data stored in HDFS and HBase. However, in the research of scientific data management methods, there are still some problems in the following aspects :(1) in the storage of massive heterogeneous scientific data, there is a lack of different storage methods for data according to different importance levels and frequent access.(2) in terms of the establishment of indexes, the practice method is too complicated based on the relationship between tables in the relational database, and the non-relational database (NoSQL) is lacking in the method of fast establishment of indexes. For this, based on scientific data mass heterogeneous characteristics, using the mongo and HDFS in storage and access to scientific data, the characteristics of the scientific method of data management, and to achieve scientific research project implementation of the scientific data storage and access, in the process of this method has the advantage of long-term preservation of scientific data, and improve the speed of fast access to scientific data.

2. Analysis of Data Management Techniques in Scientific Research Projects

The efficient management of scientific data generated during the implementation of scientific research projects needs to rely on the development of modern information technology. Traditional data management relies on the traditional database to store data, and through the connection between tables, multiple tables are associated to achieve the query and access of data, with low efficiency and slow response speed. With the development of modern information technology, the storage mode of massive heterogeneous data has changed. This paper considers using distributed file system and non-relational database to manage data.

2.1. Distributed File System(DFS)

Distributed File System (DFS) means that the physical storage resources managed by the File System are not necessarily directly connected to the local node, but are connected to the node through the computer network. The distributed file system can effectively solve the problem of data storage and management [15]. Data set from the scientific research project in the HDFS, is on the basis of the size of the default data block in the hadoop shard, thus through metadata node information management information of each data block, the size of the data block in the Hadoop2.0 version the default setting is 128 m/block, the implementation of scientific research project, can, according to the actual demand to set the size of the block of data to change the size of the block of data files. In addition, HDFS will copy data block information into multiple copies, which will be stored in different server file systems, and then the data block and copy will be stored in datanodes [16] to realize the data backup mechanism, which to some extent reduces the problem of scientific data loss.

2.2. Non-relational Database (Nosql)

Non-relational databases (NoSQL) include document databases such as mongoDB, in-memory databases such as redis, graph databases such as Neo4j, and column database Hbase on top of Hadoop. The data generated in scientific research projects include structured data, semi-structured data and unstructured data. The use of non-relational database to process scientific data with different storage structures is relatively easy compared with the traditional structured database, and the query efficiency of the database is optimized. Scientific data generated in this paper, we study the scientific research projects, including research data also includes the study of metadata information, manage the metadata information, can further manage research data, therefore, the author proposed directing a document type database to store the data of the metadata information, at the same time, through access to detailed information and view the metadata, and then obtain metadata related research data, namely the generated in the raw data of scientific research project.

3. Research on Scientific Data Management Methods

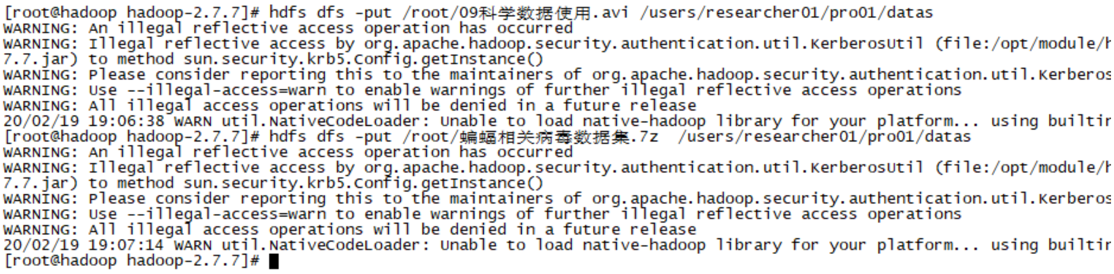
Based on the above research, this paper analyzes the characteristics of modern information technology and applies modern information technology to the field of scientific data management, so as to ensure the high availability of scientific data and realize the long-term preservation and fast retrieval of scientific data.

3.1. Management of Heterogeneous Data

At present, during the implementation of scientific research projects, data of different formats and types are generated, including structured, semi-structured and unstructured data, such as text, pictures, video and audio [14]. Based on the distributed storage of HDFS and the idea of data backup, the author stores heterogeneous scientific data in the distributed file system (HDFS). In addition, this paper will introduce the metadata information of MongoDB document database to store scientific data, so as to ensure the high availability of data and improve the speed of data access.

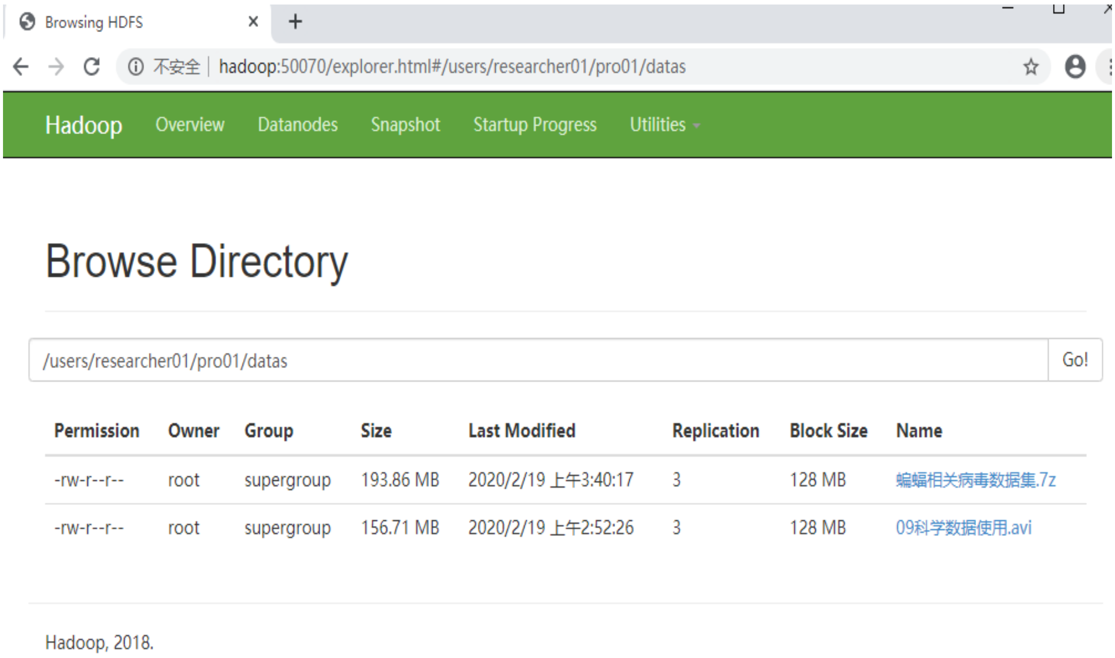
3.1.1. Distributed Storage of Scientific Data Sets

In order to facilitate the storage of data and realize distributed storage of data sets, the author USES the concept of distributed storage of HDFS and the mechanism of backup to upload the data set to HDFS for storage. The command to upload the local data set to the specified path of the specified HDFS server is: hdfs dfs -put /root/bat-related virus data set /.7z /users/researcher01/pro01/datas, among these , the command of ”-put” is to upload the file to the HDFS server, ” /root/ bat-related virus data set.7z” is the data set about bat virus on the local server /root path in the local liunx system, ” /users/researcher01/pro01/datas” is the path on the HDFS server to store the data set to be uploaded, and the data access address to be stored in the metadata information. The command to store the dataset, see Figure 1. where two dataset files are uploaded to the HDFS distributed file system from the Linux server's local path. The details of the data set uploaded to the distributed file system can be viewed on port 50070, see Figure 2.



```
[root@hadoop hadoop-2.7.7]# hdfs dfs -put /root/09科学数据使用.avi /users/researcher01/pro01/datas
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/opt/module/h
7.7.jar) to method sun.security.krb5.Config.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.authentication.util.Kerberos
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
20/02/19 19:06:38 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtir
[root@hadoop hadoop-2.7.7]# hdfs dfs -put /root/蝙蝠相关病毒数据集.7z  /users/researcher01/pro01/datas
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/opt/module/h
7.7.jar) to method sun.security.krb5.Config.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.authentication.util.Kerberos
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
20/02/19 19:07:14 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtir
[root@hadoop hadoop-2.7.7]# █
```

Figure 1: HDFS stores the dataset



Browse Directory

/users/researcher01/pro01/datas

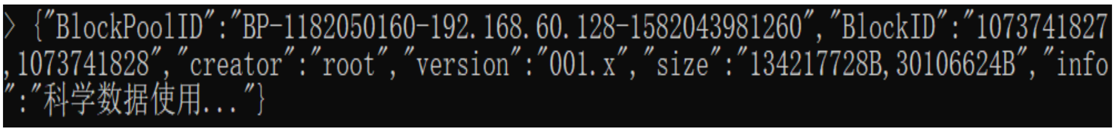
Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	root	supergroup	193.86 MB	2020/2/19 上午3:40:17	3	128 MB	蝙蝠相关病毒数据集.7z
-rw-r--r--	root	supergroup	156.71 MB	2020/2/19 上午2:52:26	3	128 MB	09科学数据使用.avi

Hadoop, 2018.

Figure 2: data set details in HDFS

3.1.2. A Dataset Describes the Storage of Information

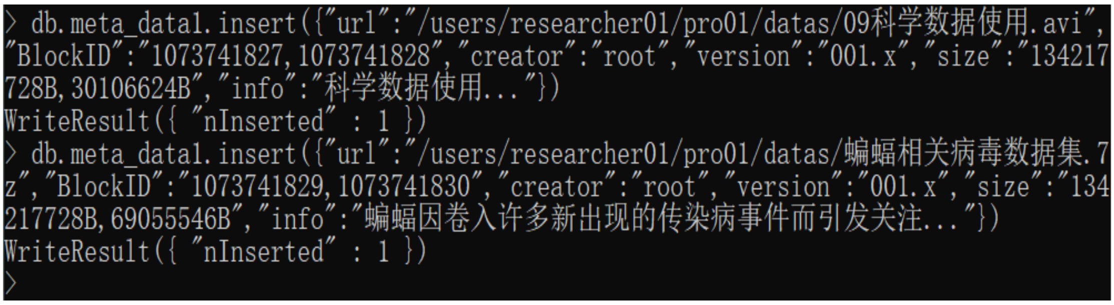
Scientific data management according to different subjects and themes, generated from the scientific research project of scientific data to classify, analysis and processing of scientific data, the form data set, and extract the detailed information of the data set is created for each dataset the only resource locator, thus ensuring the uniqueness of the data set, the data set description and uniform resource locator encapsulated into packets, form a data set of metadata. Metadata is to describe data sets of data, at the same time, through the metadata of the uniform resource locator to find the corresponding data set, including its basic content, the basic information of the data: uniform resource locator, author, version, and the size of the data, data format, data collection of creation date, the latest release date and the number of data records, data description, description, keywords, such as data, contact information, contact person, contact phone number, email, address, etc., for example, the description "scientific data use" data set of metadata information format, see Figure 3.



```
> {"BlockPoolID":"BP-1182050160-192.168.60.128-1582043981260","BlockID":"1073741827
,1073741828","creator":"root","version":"001.x","size":"134217728B,30106624B","info
":"科学数据使用..."}
```

Figure 3: metadata information

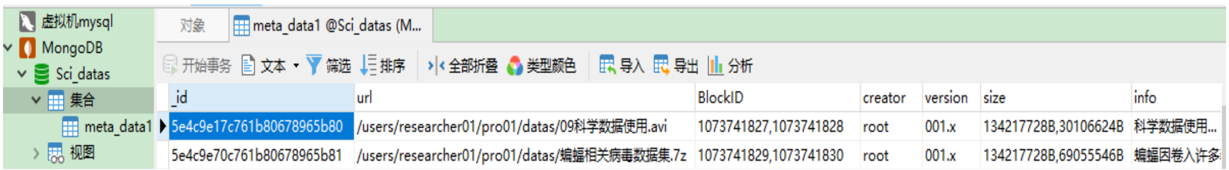
The metadata information, see Figure 3, is json-like data that can be easily stored in non-relational databases, including MongoDB databases. SQL statements that store metadata information in MongoDB are: db. meta_ data1. Insert ({“url”:” /users / researcher01 /pro01 /datas/09 use of scientific data.avi”,}), among these, ”db” represents the currently selected database, ”meta_data1” is a collection of documents created in a database, the information between ”{” and ”}” is the metadata details of a dataset. The corresponding statement of storing metadata information in MongoDB database, see Figure 4.



```
> db.meta_data1.insert({"url":"/users/researcher01/pro01/datas/09科学数据使用.avi",
"BlockID":"1073741827,1073741828","creator":"root","version":"001.x","size":"134217
728B,30106624B","info":"科学数据使用..."})
WriteResult({ "nInserted" : 1 })
> db.meta_data1.insert({"url":"/users/researcher01/pro01/datas/蝙蝠相关病毒数据集.7
z","BlockID":"1073741829,1073741830","creator":"root","version":"001.x","size":"134
217728B,69055546B","info":"蝙蝠因卷入许多新出现的传染病事件而引发关注..."})
WriteResult({ "nInserted" : 1 })
>
```

Figure 4: insert metadata

Two metadata messages are inserted, see Figure 4, among these, ”url” is the address of the actual storage location of the scientific dataset described in the metadata, through which the dataset can be accessed,”BlockID” is the ID of the data block stored by the data set in HDFS. Within HDFS, the node of nameNode can manage the information of the data block, so as to obtain the specific data block of the data set, and then integrate each data block to form the original data set. The data stored in MongoDB can be viewed by visual tools, see Figure 5. It can be seen that a unique "_id" index is automatically created for each metadata information in the MongoDB database, through which metadata information can be accessed and obtained.



_id	url	BlockID	creator	version	size	info
5e4c9e17c761b80678965b80	/users/researcher01/pro01/datas/09科学数据使用.avi	1073741827,1073741828	root	001.x	134217728B,30106624B	科学数据使用...
5e4c9e70c761b80678965b81	/users/researcher01/pro01/datas/蝙蝠相关病毒数据集.7z	1073741829,1073741830	root	001.x	134217728B,69055546B	蝙蝠因卷入许多

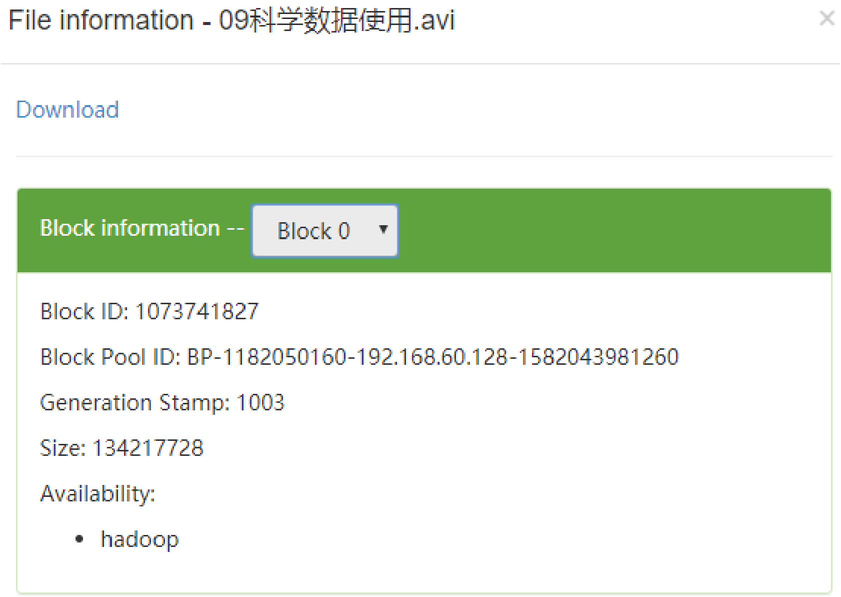
Figure 5: visualization of metadata information

3.2. Scientific Data Management Methods

Based on the above analysis, on the basis of existing research on scientific data management methods, this paper combined MongoDB document database with HDFS distributed file system to study a new method of scientific data management: unified management of distributed storage and metadata information. This method solves the problem of data loss and long-term storage to some extent, which can be analyzed from the following two aspects: HDFS distributed storage, MongoDB storage metadata information, and data access and acquisition.

3.2.1. HDFS Distributed Storage

The hadoop distributed file system storage and long-term preservation of scientific data, distributed file systems HDFS storage data sets, using the nameNode and the dataNode to manage scientific data together, first of all, the nameNode in recording the data of the stored data set piece of pool ID information, ID information of each data block, the number of data backup, data description information, etc.; Second, dataNode is specific data block node storing scientific data set, each stored in a data set to the HDFS, is on the basis of setting the size of the block of data, such as hadoop2. Default is 128 m, x shard the data set, the formation of multiple pieces of data, and to write a block of data information into the nameNode node, so as to realize the backup and distributed storage of data set, the block of data information after storage, see Figure 6, Figure 7.



File information - 09科学数据使用.avi

Download

Block information -- Block 0

Block ID: 1073741827

Block Pool ID: BP-1182050160-192.168.60.128-1582043981260

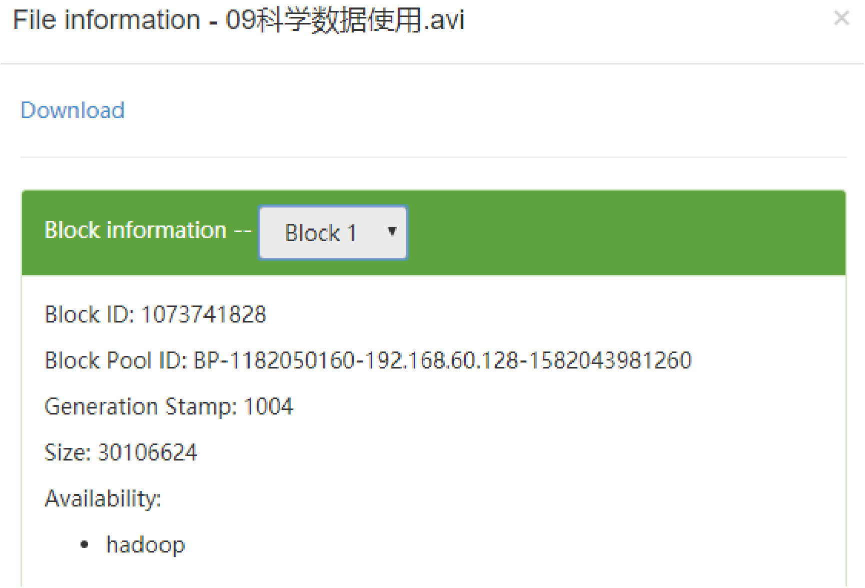
Generation Stamp: 1003

Size: 134217728

Availability:

- hadoop

Figure 6: details of data block 0



File information - 09科学数据使用.avi

Download

Block information -- Block 1

Block ID: 1073741828

Block Pool ID: BP-1182050160-192.168.60.128-1582043981260

Generation Stamp: 1004

Size: 30106624

Availability:

- hadoop

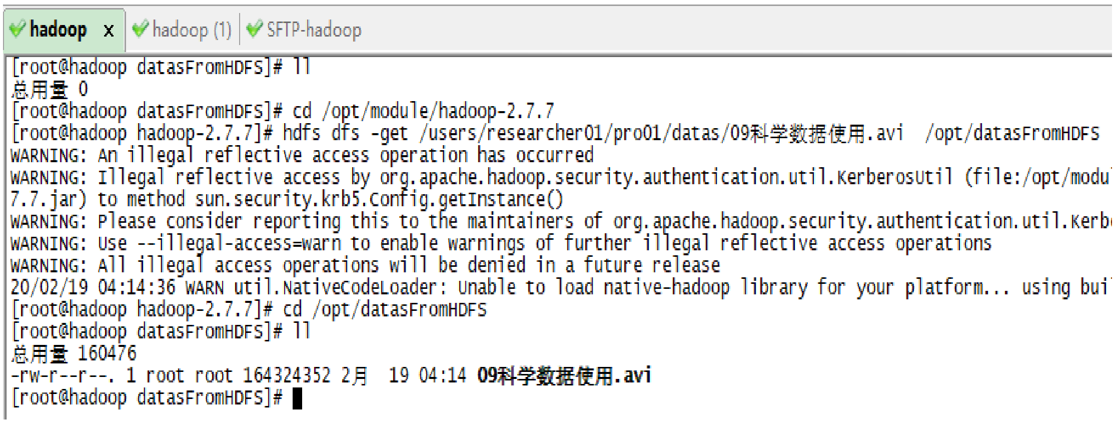
Figure 7: details of data block 1

3.2.2. MongoDb Stores Metadata

Preserved by directing a database to store data set of metadata information on HDFS, describe the metadata information of the data set is a class of the json format some information using a combination of string, the format of the information, subject to the actual demand of the scientific research project, the built-in json format, thus increasing the efficiency of data search, the researchers by accessing the metadata information in the mongo, and then screening data set of information related to scientific research project, without direct access to the stored data set in the hadoop distributed file system, this reduces the time of frequent IO calls caused by accessing disk data and increases the efficiency of data access.

3.2.3. Data Access and Retrieval

Data sets the description of the information stored in the mongo collection in the database, each record in the collection, each data set is stored at the HDFS information, including data collection of information stored in the location of the HDFS URL, so in this paper through the metadata information of corresponding value in the URL, and then get the detailed information of the data set, see Figure 8.



```
[root@hadoop datasFromHDFS]# ll
总用量 0
[root@hadoop datasFromHDFS]# cd /opt/module/hadoop-2.7.7
[root@hadoop hadoop-2.7.7]# hdfs dfs -get /users/researcher01/pro01/datas/09科学数据使用.avi  /opt/datasFromHDFS
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/opt/modu
7.7.jar) to method sun.security.krb5.Config.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.authentication.util.Kerb
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
20/02/19 04:14:36 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using bui
[root@hadoop hadoop-2.7.7]# cd /opt/datasFromHDFS
[root@hadoop datasFromHDFS]# ll
总用量 160476
-rw-r--r--. 1 root root 164324352 2月  19 04:14 09科学数据使用.avi
[root@hadoop datasFromHDFS]# █
```

Figure 8: acquisition of the dataset

According to figure 8, firstly, the files under the "/opt/datasFromHDFS" folder are empty through Linux ll command; secondly, according to the value of url key in the metadata information stored in mongoDB database" /users/researcher01/pro01/datas/09 use of scientific data.avi" , use the -get command of HDFS to download the data set stored in HDFS to the local "/opt/datasFromHDFS"; Finally, look again at the file in the Linux local server /opt/datasFromHDFS folder and see that there is a data set "09 use of scientific data.avi", which is the data set downloaded from HDFS. Thus, mongoDB and HDFS are combined to realize data storage and access.

4. Conclusion

This paper studies the scientific data management method combining the HDFS distributed file system with the non-relational database MongoDB. On the one hand, this method USES the data set generated in the HDFS storage research project, which has been analyzed and processed. On this basis, the storage address of the data set is formed.On the other hand, the method extracts the storage address and description information of the data set to form the metadata information of the data set, describes the metadata information in json-like format, and stores it in the MongoDB document database.Through metadata information stored in MongoDB, researchers browse and filter the data set related to the research topic, obtain the corresponding value of the url key of the data set, and then download the data set stored in HDFS to realize the storage and access of data.This method is applied in the field of scientific data management and provides reference for scientific data management.However, in the research process, the author only considers the characteristics of massive heterogeneous data, and has not yet studied the data with small data volume. Therefore, the next step of the author will continue to study scientific data management methods, improve scientific data management methods in the field of scientific data management, and provide reference for the subsequent scientific research on data management.

References

[1] Yuzhuo Chi, Yanfei Wang. Analysis framework of scientific data reference content oriented to scientific data management [J]. Journal of information science, 2018,37(01):43-51.

[2] Ying Xiang, Jianfei Lai, Ning Ding. Practical exploration of scientific data management service in university library -- a case study of social science data management in wuhan university [J]. Information theory and practice, 2013(12):93-97.

[3] Shaoxiong Fu, Xiaoyu Chen, Haiping Zhao, Anqi. Practice system of scientific data management in Singapore universities [J]. Library forum,2019,39(02):141-148.

[4] Wenjing Chu, Shuning Li. Research on standardized process of scientific data management in colleges and universities [J]. Intelligence theory and practice, 2019, 42(02):66-71.

[5] Hang Li. Construction of scientific research data management system of academic library based on data life cycle model [J]. Journal of library, 2016(12).

[6] Li Si, Yueliang Zeng. Research progress of scientific research data knowledge base of foreign institutions [J]. Acta sinica sinica,2017,36(08):859-870.

[7] Haitao Li, Yu Guan, Haiguang Huang. Ocean science data management and visualization platform [J]. Computer system application,2017,26(09):62-68.

[8] Lili Zhang, Liangming Wen, Lei Shi, Xiaohuan Zheng, Jianhui Li. Recent advances in scientific data management and open sharing at home and abroad [J]. Proceedings of the Chinese academy of sciences, 2018,33(08):774-782.

[9] Jingran Ma, Yuejin Zeng. Data standardization and scientific management methods [J]. China education technology & equipment,2008(24):128-130.

[10] National Science Foundation. Chapter II-Proposal Preparation Instructions.[2018-07-01]. https: //www.nsf.gov/pubs/policydocs/pappguide/nsf11001/gpg_2.jsp#dmp.

[11] Fischer M , Prabhune A , Schwarz K , et al. Advancing data management and analysis in different scientific disciplines[C]// 2017.

[12] Stillerman J, Greenwald M, Wright J . Scientific data management with navigational metadata[J]. Fusion Engineering and Design, 2018, 128:113-116.

[13] Huang X , Wang L , Yan J , et al. Towards Building a Distributed Data Management Architecture to Integrate Multi-Sources Remote Sensing Big Data[C]// The 20th International Conferences on High Performance Computing and Communications. IEEE Computer Society, 2018.

[14] Xiangbao Meng, Peng Qian. Research on data characteristics of humanities and social sciences from the perspective of data life cycle [J]. Library information knowledge,2017(01):76-88.

[15] Zhi Song. Application of distributed storage technology to optimize provincial CIMISS data service capability [J]. Meteorological science and technology, 2019, 47(3):433-438.

[16] Ting Yang, Meng Wang, Yajian Zhang, Yingjie Zhao, Haibo Pen. Energy saving optimization algorithm for differential storage of HDFS in cloud computing data center. Journal of computer science,2019,42(4):721-735.