Research on the Key Technology of Big Data Service in University Library

Chunlei Ye
Information Consultation Department
Library of Beijing University of Agriculture
Beijing, China

Abstract

With the development of information technology of university library, the mass data of the university library has the basic characteristics of Big Data. However, the current situation of the university library is the lack of distributed storage and computing model for massive data, the lack of capacity to handl the diverse data sources, including the structured, semi-structured and unstructured data, the lack of a simple, flexible application model of big data service.In order to solve the problems in the service innovation of University Libraries in China, such as the problem of distributed storage and computation of massive data, the distributed management of diverse data sources, the simple and flexible application of big data services, this paper analyzes the research contents of big data processing, Hadoop ecosystem and the demand for big data services in University Libraries, and presents a technology framework for big data service in University Libraries based Hadoop. The framework includes the distributed storage and parallel computing model of mass data, the distributed management model of diverse data sources and the model of diversified service application for university libraries. This framework takes full account of the service innovation change of University Library under the environment of big data, such as data storage and calculation, data management and service applications et al. It can solve the key technical problems of big data service of University Library in a certain extent.

Keywords: Hadoop; Big Data; hadoop; big data service ; University Libraries

I. INTRODUCTION

With the development of information technology of university library, a large number of digital resources have entered the library. The popularity of the mobile terminal allows readers to obtain knowledge without the limitations of time and space, so that the amount of user data also grew explosively. At the same time, the data source of the university library also presents the diversified characteristic. In addition to the traditional structured data resources, there are a large number of semi structured / unstructured data resources. And with the popularity of social networking applications, the speed of Internet data generated more than any previous media. So the amount of data produced by the users of the university library is larger, which forms the basis of the rapid development of big data. Therefore, regardless of the type, quantity, value or development trend, the massive data of the university library has the basic characteristics of Big Data.

As a practical position in the field of Library and Information Science, Researchers in this field have been paying close attention to the application of new information technology. At present, the domestic researchers have carried on the research on library development under the big data environment. Most of the scholars discussed the feasibility, the value and the significance, challenges and existing problems, construction path of the application of big data in university libraries[1-3]. But, it was lack of the research on the whole application of the big data technology in the university library service from the perspective of the big data technology.

Therefore, in order to give full play to the role of big data technology in promoting the service innovation of university library, this paper proposes a kind of technology framework of big data service in university libraries based on Hadoop. The key technologies involved are discussed from the perspective of practical application. This paper establishes a preliminary technical model, aiming at solving the three major problems in the service innovation of university uibraries under the big data environment. Firstly, the lack of distributed storage and computing model for massive data; secondly, the lack of capacity to handle the diverse data sources, including the structured, semi-structured and unstructured data; thirdly, the lack of a simple, flexible application model of big data service. Therefore, the focus of this paper mainly includes three aspects: the first one is to build the distributed data storage and parallel calculation model of based on HDFS and MapReduce, the second one is to build a diversified distributed data management model based on HBase to handle the diversified data source, the third one is to build the diversified service processing model of big data based on Hive and Pig.

II. RESEARCH STATUS OF BIG DATA

Big data technology is the integration of many computing technologies. From the point of view of information system, the processing of big data can be divided into infrastructure layer, system software layer, parallel algorithm and application layer. As a new distributed storage and parallel computing architecture, Hadoop can be deployed on a common platform.

Beijing Social Science Fund Project (NO: 16XCB006);

978-1-5386-2165-3/17/$31.00 ©2017 IEEE

Because Hadoop has the advantages of scalable, low economical, high efficient and reliable etc, It is widely used in the field of distributed computing, and it has gradually become the parallel processing standards of massive data in the industry and academia. As a kind of software system, Hadoop mainly includes two parts: distributed storage which is supported by HDFS and parallel computing which is supported by MapReduce.

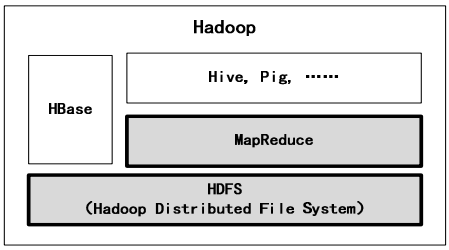

Figure 1. The basic architecture of Hadoop

Hadoop references the GFS (Google File System) [4]to achieve the HDFS(Hadoop Distributed File System), references the MapReduce[5] computing model to achieve the distributed computing framework, and references the BigTable[6] to achieve the HBase as well as other tools, such as Hive, Pig, etc. HDFS and MapReduce are two core subsystems in Hadoop.

The system software layer of big data processing mainly considers the storage management and parallel computing. Chen C proposes a new distributed storage architecture of digital data. But the main idea of the design are from the HDFS[7]. Liang J R introduced a complex storage system of big data based Hadoop[8]. At present, the development of distributed file management system based on HDFS is more mature. It can support the effective storage management of large scale data files in a scalable way.

We also need to consider the storage problems of the diversity of data structure. There is generally structured data in the traditional database. But in the era of big data, more than 80% of the data is semi-structured or unstructured. And the data structure will change a lot with the time. Massive data capacity and semi structured / unstructured data structures are a huge challenge for traditional databases. At present, scholars at home and abroad have done extensive research on the processing and application of unstructured data[9-12].

When the traditional database is difficult to adapt to the semi-structured and unstructured data, the NoSQL appears[13]. But the NoSQL generally does not provide SQL language which a large number of database application developers are familiar with. Providing the SQL query mechanism on NoSQL leads to a new technology development trend, that is required to combine the traditional SQL which is oriented to the structured data query and the NoSQL which is oriented to the semi-structured and unstructured data query. Therefore, the NewSQL appears, such as the Apache HBase. HBase has been widely used in many fields because of its distributed characteristics, massive storage and flexible data definition[14-15].

After solving the storage problem of big data, the next step is how to complete the calculation of mass data quickly and effectively. The massive data scale of big data makes it difficult to deal with the data in an acceptable time. Therefore, it is necessary to use parallel computing model and framework to support big data processing. Currently the MapReduce is the mainstream. Wang S discussed exploring a hybrid architecture to achieve the data warehouse system by using parallel databases and MapReduce[16].Li J H completed the text mining of the internet public opinion through the expansion of MapReduce[17]. He S used the MapReduce to achieve personalized service based on user logs[18].

The parallel algorithm layer of big data considers how to design the parallel algorithms for the analysis and mining of the big data. Various types of mining algorithms can be achieved by writing the MapReduce program. However, the requirements are relatively high for developers. They need to write complex MapReduce program based Java. Hive is a data warehouse system which can be used for data analysis and data mining. It uses SQL-like language to describe data processing logicly, and avoids developers writing the complex MapReduce program. Wu X Y completed the data mining and analysis through calling for Mahout in the Hive[19].

Pig is a platform for handling the massive data sets which Yahoo contributes to Apache. Compared with MapReduce, Pig provides a higher level for the processing of massive data sets. The use of Pig can simplify the development of MapReduce tasks, improve the convenience of data processing of Hadoop cluster which controls with thousands of machines. Pig can be seen as a kind of client software of Hadoop, it can be directly connected to the Hadoop cluster for data analysis. Pig is particularly convenient for users who are not familiar with Java, it uses a relatively simple SQL language oriented data flow, which is called Pig Latin. However, there is little research on the further application of Pig.

Based on the above analysis it can be seen that the Hadoop provides diversified, flexible and scalable members to complete the requirements of big data processing. Therefore, Chen J R proposed that Hadoop ecosystem would be the first choice for SMEs to solve the big data problems[20]. Zhang H introduced the Wenjin search system which was developed independently by National Libraries and software developers, Hadoop and NoSQL were applied into this system[21]. However, there has been little application of Hadoop ecosystem to the overall service of big data in university libraries presently. Therefore, this paper proposes a technology framework for the overall

service of big data in university libraries based on Hadoop, which combines the technology content of big data, the Hadoop ecosystem structure and the information service demand of big data in university libraries. The feasibility of the framework is verified by simulation experiments in a single node pseudo distributed environment.

III. THE FRAMEWORK AND KEY TECHNOLOGIES

A. concepts and framework

Generally, the processing of big data refers to the use of the techniques and tools of big data to process the data, and the extraction and integration of heterogeneous data sources widely , and the storage of data according to certain standard unified, and the data mining by use of the data analysis and calculation tools. It finally show the result to the user the right way. Therefore, this paper gives the definition of big data service in university library, that is the storage, processing and analysis of big data, providing users with data display, and a variety of auxiliary decision-making in order to find the potential value of big data.

In this paper, we research the content of big data technology, the Hadoop ecosystem architecture and the requirement of big data service in university library, this paper puts forward a Hadoop based framework for big data service adequately. And we propose a framework for big data service in University Libraries based on Hadoop ecosystem, as shown in Figure 2.

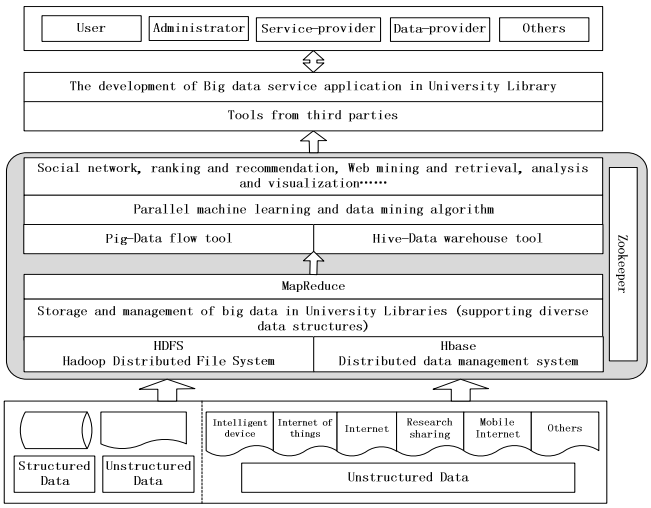

Figure 2. The framework for big data service in University Libraries based on Hadoop ecosystem

B. The key technology

This framework involves three key technologies. The first one is the distributed storage and parallel computing model, the second one is distributed management model for diverse data sources, and the third one is the diversified service application model.

- **The distributed storage and parallel computing model**

HDFS provides a scalable, highly reliable, highly available distributed storage management system for large scale data. HDFS has the ability to store data on a large scale. It can not only store a single file of GB level to TB level, but also support the storage of up to tens of millions of files in a file system. HDFS provides high data access bandwidth with multi node concurrent access mode, and can extend the scale of bandwidth to all nodes in the cluster. In the design concept of HDFS, the hardware failure is regarded as a normal. Therefore, the design of HDFS ensure that the system can automatically recover from the failure quickly, and ensure that the data is not lost. The HDFS has been optimized for sequential reads to support fast sequential reads of large amounts of data. HDFS uses multiple copies (default 3 copies) data redundancy storage mechanism, and provides an effective data error detection and data recovery mechanism, greatly improves the reliability of data storage.

According to the characteristics of the HDFS, this paper establishes the HDFS structure of big data in university library as shown in Figure 3.

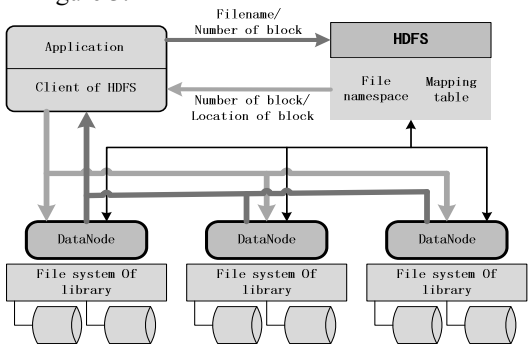

Figure 3. The basic structure of HDFS in University Library

The HDFS distributed storage system includes a master node which is called NameNode, and a set of slave nodes which are called DataNode. The big data files collecting the data of university library can be stored in the local Linux file system of several DataNodes.

When our application needs to access the university library file system, HDFS helps us to complete the task with the following steps:

Step 1: the user's application sends the file name to NameNode through the HDFS client program.

Step 2: After the NameNode receives the file name, it will retrieves the corresponding data block in the HDFS directory, and then finds the DataNode address of data block according to

the data block information. Finally, the NameNode will sent the address to the client.

Step 3: After the client receives these addresses of DataNode, it will perform the data transfer operations in parallel with these DataNode. At the same time, the relevant log of all the operations will be committed to NameNode.

However, Hadoop provides abundant Java API to support the development of Hadoop application program. The API associated with HDFS is located in the package of 'org.apache.hadoop.fs'.

The process of accessing HDFS using API programming typically includes the following steps:

Step 1: creating HDFS configuration information object Configuration through below command:

Configuration conf=new Configuration ();

Step 2: creating FileSystem object based on Configuration object through below command:

FileSystem hdfs=FileSystem.get (conf);

Step 3: calling the corresponding method of FileSystem file operations through below command :

FileStatus files[]=hdfs.listStatus (path).

In the Hadoop ecosystem, MapReduce adopts the thought of 'divide and rule'. It distributes the processing tasks of the large scale dataset to the multiple slave nodes (Slaver) which are controlled by a master node (Master). The final result is obtained by integrating the intermediate results from each Slaver. The MapReduce contains two core abstract methods, which is Map() and Reduce().The method of Map() is responsible for decomposing the task into multiple subtasks, and the method of Reduce () is responsible for summarizing the processing results of these subtasks. The MapReduce also provides an Combiner object that is specifically responsible for the optimization of data transfer for intermediate results.

We can write a variety of MapReduce applications, and compile it into the Jar package, which is carried on the Hadoop system, such as below command:

```
[root@winstar ~]# hadoop jar InvertedIndex.jar wcin wcmout
```

- **Distributed management model for diverse data sources**

HBase is a distributed scalable NoSQL database based on HDFS. It provides the ability to read, write and random access to structured, semi-structured and unstructured data. HBase also provides a triple- dimension data management model based on row, column, and timestamp. The HDFS stores data in the form of key-value pairs: {row key, column family, column name, timestamp}->value. For example, a key-value pairs to store the data of university library showing the user accessing resources can be expressed as: {key3, userInfo, dataSource, t2}->'http://www.cnki.net'. A mapping table is stored in each row and each column family. Each column does not need static definite, it can be added or deleted dynamically. For university library user, they access to or use of the library in the diversified form. Therefore, the structure of user data is also diversified. HBase can provide accurate data preservation through the flexible way. At the same time, the number of row per table in HBase can be as many as several billion or more, each row can have up to millions of columns. And this storage capacity does not require special hardware, ordinary server clusters can be competent. Therefore, it is feasible for the library to store the diversified and massive data based on HBase solution.

As an independent Hadoop subproject, HBase provides a complete set of API supporting to access and operate the HBase database through Java programming. The Java application of client that accesses the HBase database uses the API in 'org.apache.hadoop.hbase' and 'org.apache.hadoop.hbase.client'. Although HBase supports MapReduce programming, but the general client program does not need to run on the Hadoop. HBase applications only need to achieve through the Java project. All operations on the database can be encapsulated in the Java application.

- **The diversified service application model**

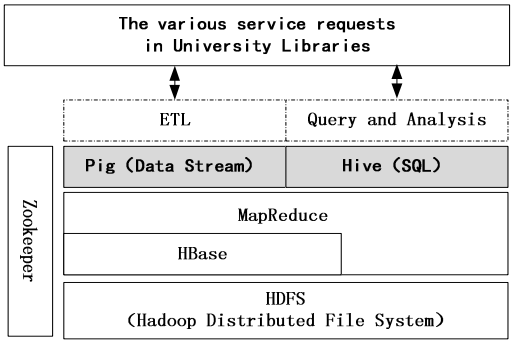

Figure 4. The big data service processing framework based on Pig and Hive

Hive is a tool of data warehouse based on Hadoop, it was first used for processing and analysis of large amount of user log data by Facebook. It provides the programming interface of HiveQL(Hive Query Language), and it has the functions of data extraction, storage management and query analysis et al. Hive not only provides the default Hive Shell services, but also provides JDBC driver, which can directly be integrated with Java applications.

The Java applications of Hive accessing Hive JDBC services are almost similar with general JDBC applications, which includes using the DriverManager to load a specific Hive JDBC driver, building a session connection based on Hive server, creating a record set of HiveQL query ,

submitting query and receiving record set, and extracting data from the record set.

In the JDBC applications, Hive can achieve flexible analysis and processing of data by using HiveQL, which is similar with the SQL.

```
<terminated> HiveJdbcTest [Java Application] /usr/jdk1.7.0_71/bin/java
SELECT * FROM student
11      John    100     class1
20      Marry   93      class2
31      Peter   87      class3
24      Smith   98      class2
15      Linda   85      class1
26      Jane    86      class2
select id, name, score from student where class='class2'
20      Marry   93
24      Smith   98
26      Jane    86
select class, count(*) from student group by class
class1  2
class2  3
class3  1
```

Pig Latin is an language oriented dataflow, and it is a specific script language of Pig that is similar to the SQL. The script language provides the function of sorting, filtering, summing, grouping, and association et al. At the same time, it allows users to customize some functions to meet some special needs of data processing. When it is needed to deal with massive data, we write the script program with Pig Latin first of all, and then the script program is executed in Pig. The Pig Latin program will be compiled for the MapReduce program, and be uploaded to the cluster to run.

We write a Pig Latin script program which is named "WordCount.pig", the script code is shown as follows:

```
1  inp = load '$inputdir/*' as (line:chararray);
2  words = foreach inp generate flatten(TOKENIZE(line)) as word;
3  grpd = group words by word;
4  wcnt = foreach grpd generate group, COUNT(words);
5  DUMP wcnt;
6
```

Therefore, through the above analysis we can see that in the current status of massive data resources, limited fund foundation, diversified of service demand of university library, the framework of big data service is proposed in this paper will provides sufficient feasibility for big data service of university library. It can help to fully dig out the potential value of massive data resources, which will improve the service innovation of university library.

IV. Conclusion

In order to solve the problems in the service innovation of University Libraries in China, such as the problem of distributed storage and computation of massive data, the distributed management of diverse data sources, the simple and flexible application of big data services, this paper analyzes the research contents of big data processing, Hadoop ecosystem and the demand for big data services in University Libraries, and proposes a technology framework for big data service in University Libraries based Hadoop. The framework includes the distributed storage and parallel computing model of mass data aiming to store the mass data of University library, the distributed management model of diverse data sources aiming to carry out effective management of diversified data resources of University library, and the model of diversified service application aiming to provide a simple, flexible and diversified big data service for University library. This framework takes full account of the service innovation change of University library under the environment of big data, such as data storage and calculation, data management and service applications et al. It can solve the key technical problems of big data service of University library in a certain extent.

However, this paper only proposed a feasible framework, and the concrete application research needs to be researched further.

Acknowledgment

This paper is sponsored by Beijing Social Science Fund Project. The number of project is 16XCB006. And it is also sponsored by CALIS National Agricultural Literature Information Center Project. Then number of project is 2016008. Thanks for the sponsor.

References

[1] Su. X. N, “Opportunities and Challenges Faced by Digital Libraries in the Era of Big Data,” Journal of Library Science in China, vol 06, pp 4-12, 2015.

[2] Chen. C. F, Qian. O, and Dai Y Z, “Study on the Construction of Digital Library in the Age of Big Data,” Library and Information Service, vol 07, pp 40-45, 2014.

[3] Cheng. J. J, “Research on the content and strategy of Library Service Innovation in the era of big data,” Information Studies: Theory & Application, vol 03, pp 57-62, 2016.

[4] Ghemawat. S, Gobioff. H, Leung. ST, “The Google File System,” Proceedings of the 19th ACM Symposium on Operating System Principles(SOSP2003）, New York, USA, pp 29-43 ,2003.

[5] Dean. J, Ghemawat. S, “MapReduce: Simplified data processing on large clusters,” Proceeding of the Conference on Operating System Design and Implementation(OSDI), San Francisco, USA, pp 137-150, 2004.

[6] Chang. F, Dean. J, Ghemawat. S, et al, “Bigtable: A distributed storage system for structured data,” Proceeding of the 7th Symposium on Operating Systems Design and Implementation(OSDI). Seattle, USA, pp 205-218, 2006.

[7] Chen. C, “A Distributed Big Data Storage Architectures for Digital Library Based on New Storages,”Journal of Modern Information, vol 01,pp 100-103,2015.

[8] Liang. J. R,”Study of the Compound Big Data Storage System for Library Based on Hadoop,” Journal of Modern Information,vol 02,pp 63-67, 2017.

[9] Ferrucci. D, Lally. A, “UIMA: an architectural approach to unstructured information processing in the corporate research environment,” Natural Language Engineering, vol 10, pp 327-348, 2004.

[10] Doan. A, Naughton J F, Baid A, et al, “The case for a structured approach to managing unstructured data,” Proceedings of the 4th Biennial Conference on Innovative Data Systems Research. United States: CIDR, pp 1-10, 2009.

[11] Han. J, E. H. H, Song. M. N, et al, “Model for unstructured data based on subject behavior,” Computer Engineering and Design, vol 03, pp 904-908, 2013.

[12] Bai. R. J, Leng. F. H, “Research on the integration of scientific data in the era of big data,” Information Studies: Theory & Application, vol 01, pp 94-99,2014.

[13] Shen. D. R, Yu. G, Wang. D. R, et al,”Survey on NoSQL for Management of Big Data,”Journal of Software, vol 08, pp 1786-1803, 2013.

[14] Wang. Y, Tao. Y, Yuan. J, et al, “Approach to Process Smart Grid Time-Serial Big Data Based on HBase,” Journal of System Simulation, vol 03,pp 559-568, 2016.

[15] Xu. A. P, Wang. B, Xu. W. P, “Research on associated query of monitoring video big data base on spatio-temporal characteristics in HBase,” Application Research of Computers, vol 05,pp 1-7, 2017.

[16] Wang. S, Wang. H. J, Qin. X. P, et al, “Architecting Big Data: Challenges, Studies and Forecasts,”Chinese Journal of Computers, vol 10,pp 1741-1752, 2011.

[17] Li. J. H, He. Y. S, Xiong. Q, “On Text Mining of Network Public Opinion Based on Big Data Technology,”Journal of Intelligence, vol 10, pp 1-6, 2014.

[18] He. S, Feng. X. L, Wu. Q. H, et al,”Research on Personalized Services of Library Based on User Behavior Modeling and Big Data Mining,” Library and Information Service, vol 61,pp 40-46, 2017.

[19] Wu. X. Y, Ming. J. R,”Research on the Big Data Management Model Based on Data Mining,” Information Science, vol 33, pp 131-134, 2015.

[20] Chen. J. R, Le. J. J, “Reviewing the big data solution based on Hadoop ecosystem,” Computer Engineering & Science, vol 10, pp 25-35, 2013.

[21] Zhang. H,“Application of Big Data Technology in Building Resource Discovery Platform: Taking Wenjin Retrieval System as an Example,” Digital Library Forum, vol 01, pp 61-67, 2016.