Cloud Service Research on Big Data Computing in Colleges

Hongkai Lin
Wuhan Business University, Wuhan, Hubei, 430056
linhk@wbu.edu.cn

Keywords: Big Data; Hadoop; RCFile; Complementary clustering index; Cloud Service.
Abstract. With the continuous development and application of information technology, the continuous construction of digital campus and intelligent campus project in colleges, the number of data stored in the college campus environment has increased dramatically, and a big data environment in a college campus has initially come into being. There are many kinds of big data in college campus, which can reflect huge value by means of information acquisition and data mining. In this paper, based on the research of Hadoop, cloud services are put forward, which can provide reliable data support for college teaching and management. It is helpful for the collection, storage, management and application of the big data of college campus in the information age, which is one of the important researches for educational and scientific researchers.

Introduction

In recent years, the development of information technology in colleges, as part of the development of colleges, is attracting more and more attention from college management. The digital resources of colleges are exponentially increasing, creating a lot of business data. These data come from various application systems within the campus, whose features are massive, complex, diverse and heterogeneous etc. Therefore, a lot of information islands have been formed. How to store and manage these massive data, break through these information isolated islands, improve the data quality, exchange the real-time data and explore the data value, is a hot topic in the research of information management for college. The cloud platform based on big data computing virtualized the hardware resources and network resources to achieve powerful storage and computing power, and realized the unified management of data and reduced the management cost, improved the quality of service, and the reliability and scalability of data processing[1].

Current situation of big data environment in Colleges.

Currently, the big data environment of college campus has initially come into being, and various kinds of campus data are widely distributed, complex in types and huge in data volume. Taking college students as an example, the big data of college students can be divided into two major categories, one is the big academic data, which includes the data of the course selection of the college students, the attendance data of each course, the daily test and test results of all the subjects, and the other is the big data of life, including the access card data, the library reading data, supermarket consumption data, even including students' WeChat, micro-blog etc. Large scale data sets and multilevel data quality challenge the storage, analysis, research and application of college campus data. The traditional management ideas and methods are gradually unable to meet the increasing demand for data processing[2]. The big data system of College Campus Based on Hadoop is designed and constructed to collect, store and deal with many kinds of data in college campus, and to provide data support for college teaching and management through data mining and data analysis. It is not a risk identification and early warning function for college students' study and life[3].

Cloud system design of big data in colleges based on Hadoop.

Hadoop. Hadoop, developed by the Apache foundation of Open Source Group, is a basic framework of distributed system, which it can run applications on multiple hardware clusters, and

form a parallel distributed system with high performance, high reliability, high scalability and low cost. Hadoop consists of a variety of elements, whose three core technologies include distributed file systems which are used to store files in cluster hardware; programming models which are used to deal with a variety of data sets; and distributed databases. With the distributed technology of Hadoop, the college campus big data system is designed and constructed, which can realize the storage, analysis and efficient processing of a large number of data of college students, and further promote the process of the construction of the college's intelligent campus[4].

Analysis of cloud service architecture. Cloud service is a new generation of data analysis and mining platform based on Hadoop. It integrates a variety of data mining algorithms based on Hadoop, which can support the data mining of government organizations and enterprises. The so-called Hadoop, which is a platform that can be provided by a stable and reliable interface and data service, can implement the MAP/Reduce algorithm and can divide the text into several units that can be repeatedly executed. In the whole platform, MAP/Reduce algorithm, distributed file system (HDFS) and distributed column storage database (HBase) will always run through all the time. By using these algorithms and database structure, the system can access to large amounts of data at high transmission rate and realize the splitting access of text data[5]. In addition, Hadoop platform can also achieve data decomposition, and complete a large amount of data analysis and processing. From the view of cloud service structure, the platform is mainly composed of several modules, such as user management module, data management module, task management module and result display module. Using a distributed file system and a distributed and storage database by column, the platform will be able to store a large amount of data and ensure that users can access to the data quickly. Using the Web interface, the platform can show the data mining results to users. Using the user management module, the platform can manage the authority of the users, so that users can only access and manage the data within authority, and use the corresponding data mining functions. Using task management module, users can conduct mining tasks and monitor the progress of tasks[6]. And each module of the platform needs to be interfaced with the backstage mining system to realize all kinds of functions, and finally use the graphics to display the results of the algorithm analysis.

Realization of data mining. From the point of view of data mining, cloud service is a parallel data mining system, which contains more than 40 data mining algorithms, and can use a variety of algorithms to complete the pre processing operation of data. In this system, there are text processing and modeling system, information collection system, data mining system and foreground interface. In the process of system operation, the information collection system will store the data into the distributed file system, and then store the description information of the webpages into the distributed storage database by column. And information collection needs parallel crawler as the main body based on MAP, and it can provide support for various acquisition models. In the text prediction and modeling system, it thus includes many kinds of text mining preprocessing modules, such as Chinese word segmentation model modeling, text feature extraction etc. Using the data mining system, many data mining algorithms, such as emotion analysis algorithm, association analysis algorithm and abstract extraction algorithm, can be used to complete the mining of massive data. With MAP, these algorithms can run independently. In addition, in parallel data mining system, it also contains data mining algorithm based on memory computing framework Spark, which can provide users with open services[7].

Research on computing technology of big data.

Line and column hybrid data storage technology. Hive is a data warehouse tool based on Hadoop. In Hive, the best part of the operations need to store and read the data, so the data storage format and access mode of Hive will greatly affect the efficiency of Hive. In the traditional model, the data is managed by using the file format of SequenceFile. The data is stored and read by the line storage mode. The line storage can only access to the data according to the mode of the saving by line and taking by line. When it needs to read a column, it needs to take out all the data first and then

extract the data of a column, whose efficiency is quite low. The storage technology by column researches a storage mode that can access to the data in two ways of by line and column, to improve the efficiency of accessing to data for Hive. Column storage structure is a data table storage format that stores record set in order of attribute values[8]. The storage structure can read the attribute data in the record according to the query, and support data compression mechanism and query execution method with the column data as the unit. Given that the attributes of queries in data warehouse applications are often less than all attributes of data tables, column storage structure can effectively improve the efficiency of query processing. If there is a data table that is stored in the traditional binary line storage technology, it will store the data in the data table line by line, and if it stores in a line and column hybrid storage format, it first divides the records into different line groups, and then slices the lines with column and store all the data in a line group in sequence of column.. In each group, all the metadata values are stored as a whole, and then the data values of each column in the group are sequentially stored. In addition, line and column hybrid storage uses compression technology column by column, that is, the data of each column in each line group is compressed individually. The advantage of this technology is that only need to decompress the columns that the query needs. The aim is to get the advantage of data compression and column storage technology simultaneously. We test the line and column hybrid storage in the Hadoop and Hive matching version and the system test uses real data loading. The test shows that the line and column hybrid storage can save 25% of the storage space compared with the original SequenceFile format, and the data loading performance is also raised by 30%. Line storage technology has become a de facto standard for data storage structures in distributed offline data analysis systems such as Apache Hive.

Complementary clustering index technology. With the increasing amount of data in network applications, in order to meet the requirements of high read and write performance, low memory hog and high reliability. HBase is applied to the database system of uni-dimensional interval query in the background of massive data,which it can position the data gradually according to related information. At the same time, it can also meet the range queries of high throughput on the primary keys. Technicians need to make the ordering for them according to the different attributes of data to support interval queries in multiple dimensions. However, HBase's technical defects make it impossible to realize its indexing method through high-speed query and low memory hog. On the basis of HBase, technicians apply complementary clustering index method and system to meet the requirements of multidimensional interval retrieval[9]. CCIndex index table makes the ordering for the value of the attributes to be searched and realizes the storage of all the data in the index table, ensures the directness of the data obtained during the multidimensional interval query, which is conducive to improving the query efficiency and the quality of the query. At the same time, it can shield the bottom data backup of HBase, use the saved data in the index table to recover the data of the record set, greatly improving the speed of the data query. The CCIndex method includes the following three aspects:

Data organization: CCIndex method organizes the copy of the backup data as a complementary cluster index table with multiple complementary and verification, which it uses the efficient continuous scan on the index table instead of the random read on the original table, thus greatly improving the performance of multidimensional interval query.

Query processing and Optimization: CCIndex method first converts query string to query plan tree, and then make a simple optimization of eliminating the repeated and combination for the query statements and then the size of the result set of the subquery is estimated based on the slice information of HBase and finally, the minimum subquery is selected to execute the query process on the corresponding cluster index table.

Data recovery: CCIndex saves the corresponding relationship between the main keys of each index table in the complementary checklist and carry out the incremental recovery of data through the complementary cluster index table and complementary checklist, which ensures the same recoverability as the data is recovered by the data copy, while adding only a small amount of memory hog[10].

Conclusions

With the further application of information technology in college campus, the continuous construction of intelligent campus engineering, the continuous development of college campus teaching and management ideology, the big data environment of college campus will be more perfected. The establishment of big data computing system based on Hadoop technology can make the important data that have been neglected originally to be a treasure. Through scientific and reasonable big data management and big data mining, it can provide the data analysis basis for college teaching management to pay attention to students' study and life, and also provide the reliable data support for colleges to formulate teaching management policies. It is foreseeable that the emergence of big data computation will spawn much more, better and more public oriented cloud service applications, and at the same time promote the development of new technology.

Acknowledgements

This work was financially supported by Scientific Research Project of Wuhan Business University : Research on the construction of university big data computing platform based on Hadoop （2016KY039）.

References

[1] Bakia, M. , Murphy, R. , Anderson, K. , & Estrella-Trinidad, G. International experiences with technology in education: Final report[R]. U. S. Department of Education, Office of Educational Technology.(2014), 12(5), p11-25.

[2] Chang, T. W. , Hsu, J. M. , &Yu, P. T. (2016). A comparison of single-and environment in programming language；Cognitive loads and learning effects[J]. Educational Technology & Society．No S2(2016), p188-200.

[3] Hequn Wang.China Mediatech. Vol.25 (2013),p60-64.(In Chinese)

[4] Razak, S. F. A. Cloud computing in Malaysia universities[C]. Innovative Technologies in Intelligent Systems and Industrial Applications (CITISIA 2009) (2009) ,p101-106.

[5] Dongxing Jiang.et al: Journal of East China Normal University. No S1(2015),p.119-125. (In Chinese)

[6] Qing-bin Sang: Journal of Nantong Textile Vocational Technology College.Vol.13(2013), p.84-87. (In Chinese)

[7] Mayer-Schönberger V, Cukier K. Big data: A revolution that will transform how we live, work, and think. Houghton Mifflin Harcourt（2013）

[8] Arasu A,Chaudhuri S,Chen Z,et al:IEEE Data Engineering Bulletin,35-2(2017),P.14-23

[9] Labrinides A,Jagadish HV:PVLDB,5-12(2014),P.2032-2033

[10] White T. Hadoop: The definitive guide. O'Reilly Media, Inc. (2017)