Data Analysis and Parallel Database Construction of Cloud Platform

Yan Wei

Information Centre of HuaiHai Institute of Technology Lianyungang, Jiangsu, 222000, China

Keywords: Cloud Platform, Data Analysis, Parallel Database, Construction, Hadoop

Abstract: With the arrival of the era of big data, people have higher and higher requirements for data processing. Clustering analysis is the key to data processing, which can facilitate people to describe and understand data objects, and collect the intrinsic information of large data hidden in cloud platforms. It is urgent to study a reasonable parallel clustering method for large data in cloud platforms. The Internet has taken root in people's lives and spread to every corner. The development of scientific research informationization makes people deal with things more and more conveniently and synchronization more and more timely. However, the emergence of new things will always bring new problems and usher in a new pattern. When we use the Internet for information transmission and interaction and it will produce a lot of data information. However, with the rapid growth of data storage and transmission, the existing traditional computer mechanisms and technologies can no longer meet the current computing needs, which prompt people to seek new technologies to process and analyze large amounts of data and extract potential information and value from data.

1. Introduction

With the rapid growth of power grid scale and the increasing complexity of power grid structure, the deep integration of information technology and electricity for production, online monitoring of intelligent primary power equipment and conventional power equipment has been greatly developed and become a trend, monitoring data has become increasingly huge, and monitoring data acquired and transmitted in equipment has increased geometrically. On-line monitoring system of power equipment is facing enormous technical challenges in data storage, query and data analysis. How to store, access and analyze large data of power equipment monitoring efficiently and reliably is an important research topic in the field of power information processing and large data processing.

In recent years, Internet information technology has been widely used. For example, many industries, such as government units, hospital institutions, major commercial departments, catering industry and educational institutions, have more opportunities to exchange information with the outside world by means of Internet technology in a more convenient and efficient way. After a long period of continuous growth, a large number of data stored in different forms are continuously accumulated in the database. The scale of the database is increasing, the data forms are various, and the emergence of complex data. These data information will play a very meaningful reference value in the decision-making process. However, the traditional database system has been unable to achieve the requirement of obtaining deeper information in a large number of data, even when faced with massive data, it cannot be processed, and it is more difficult for people to discover and mine the hidden knowledge by intuition and experience.

Data mining algorithm is a technical means to process data, which conforms to the development trend of the times and the needs of the market. The process of data mining is essentially to obtain effective information and discover potential knowledge through data analysis, and to analyze and summarize the effective information to form a knowledge system, and feedback to relevant departments to help them make correct decisions. In the process of data mining, classification algorithm is used, which is a very basic but important technical means.

As one of the most important research directions, data mining has attracted much attention both in computer science and Internet. Due to the improvement of computer technology, the vigorous

development of Internet industry and the wide application of mobile data network, all industries have entered the road of Internet development. Data mining technology is always needed to analyze and process the data generated by information exchange on the network to obtain valuable information. Some well-known network companies, such as Facebook and Amazon, can produce hundreds of TB or even PB-level data every day. The value contained in the large amount of data will be the unique resources of these enterprises. However, the increasing amount of data also increases the load of the computing system. At present, facing the problems of single computer system in data processing, many enterprises are eager to find new technical means to better promote the development of enterprises. More enterprises focus on choosing higher performance distributed computer systems and parallel mining algorithms to improve computing capacity and promote the development of enterprises through parallel technology. Hadoop is a distributed system infrastructure developed by Apache Foundation. It has the advantages of cross-platform and high fault tolerance. Users can directly use Hadoop to organize and build their own distributed cluster platform on low-cost PC without paying attention to the design concept and detailed implementation steps of the distributed bottom layer. Thus, it is easy to use the large-scale storage and management data of the well-organized machine cluster to record the data set of parallel statistical calculation, and to accomplish the distributed task well.

Hadoop includes a distributed parallel computing framework Map Reduce, a distributed file system HDFS, as well as Pig, Hive and many other sub-projects. At present, many enterprises have built their own Hadoop distributed platform to facilitate efficient data mining.

2. The Proposed Methodology

2.1 Cloud Computing and Large Data Processing Technology.

The relationship between big data and cloud computing is as inseparable as the positive and negative sides of a coin. Big data cannot be processed by a single computer. Distributed computing architecture must be adopted, relying on distributed processing, cloud storage and virtualization technology of cloud computing.

Alibaba, the largest e-commerce company in China, has many leading websites and businesses including Alibaba, Taobao, Alipay and so on. It has very large transaction data and access data every day. Taobao is one of the first companies in China to adopt Hadoop technology for data processing. Since 2008, Taobao has been actively engaged in the research of "cloud ladder" which is the data processing platform of Hadoop technology, and has established the largest Hadoop cluster data processing platform for processing and analyzing all Taobao data. As the third-party payment platform, Alipay also occupies the leading position in China. The high reliability and high timeliness of big data processing is crucial for it. At the same time, many large domestic enterprises, such as China Mobile and Huawei, also actively participate in the improvement and application of Hadoop technology. For example, Typhoon, a cloud computing platform independently developed by Tencent, Huawei's team implemented a complex event processing component to handle complex logic judgment and high real-time requirements analysis business, and China Mobile established a cloud platform based on Hadoop technology to provide users with better and more convenient Internet business and services.

Class algorithms are more and more applied in the business field, and they are also an indispensable part of data mining research. The significance of classification is to classify things according to data. It uses classification function or model to classify the data in database into the category with the highest similarity according to their respective attributes. Firstly, the training sample data consisting of one database record is input to construct the training model. Each database discipline corresponds to a feature vector containing several attributes and a specific class label. According to different classification rules, data classification technology can be divided into statistical method, machine learning method, neural network method, genetic algorithm and so on. Among the classification methods introduced above, the decision tree method in machine learning

method has the advantages of fast speed, simple structure and high classification accuracy, which attracts much attention in the field of data mining.

Decision tree classifier generates judgment rules for decision attributes by learning data sets. Its implementation process can be understood as the process of recursive segmentation of sample space. In the decision tree algorithm, the attribute set of each data set is represented by a node of a tree. The decision tree contains all nodes from the root node to the leaf node. The branch of each node represents the attribute value of an attribute. The leaf node of the tree represents the result of the decision, that is, the category of the attribute belongs to.

2.2 Data Analysis Methods.

Because the training sample set may be noisy or the selected feature is not suitable, the decision tree generated is not simple and compact. Therefore, it is necessary to prune the decision tree, that is, by reducing some branches and leaves, the generalization ability of the model is stronger, to improve the classification accuracy of the decision tree. According to the different time points of pruning, tree pruning can be divided into synchronous pruning and lagging pruning.

Synchronized pruning is a pruning process that occurs in the process of decision tree generation. Continuously determine whether the node needs to further partition or split the subset of training samples. If not, determine the branching stop, and record the current node as a leaf node. When judging whether further division is needed, information gain method or statistical methods such as importance detection can be used to determine whether further division is needed.

Delayed pruning is a pruning process that occurs after the generation of decision tree. Delayed pruning occurs after the construction of decision tree, so it has nothing to do with the construction of decision tree. The classification algorithm generates decision tree completely according to the training data set. A complete decision tree without pruning is constructed, and a pruned decision tree is generated by delaying pruning operation. Hysteretic pruning is a process of pruning from the bottom leaf node up to the root node. The most commonly used pruning algorithm in delayed pruning algorithm is based on cost complexity. It calculates the expected error rate of each non-leaf node in the tree after pruning, and calculates the expected error rate of the node when it is not pruned according to the error rate and weight of each branch, to decide whether pruning is necessary or not.

Data itself contains a variety of the information, and a large number of data contains hidden information and value that small data sets cannot extract. With the rapid growth of data volume, traditional computer algorithms are gradually ineffective, and large data platforms emerge as the times require. The earliest big data platform is now widely used Hadoop platform. It can process data of PB level and above, and solve the problem of large data storage management and distributed computing on general commercial machines by using distributed service node group.

HDFS distributed file system has the characteristics of high fault tolerance, high concurrency, high availability and scalability. It also implements data storage management on local servers. Unlike data storage in stand-alone mode, HDFS uses several local computers to establish connections between machine service points by means of local network, and organizes a complete distributed file management system that can provide services for high-level large data processing applications. It is unrealistic to rely solely on centralized physical server storage for large-scale GB-to-TB data, whether in terms of storage capacity or transmission speed. Large data storage needs to use multiple server nodes, dozens, hundreds or more. Distributed file system must be used to manage the data storage of these nodes. HDFS, as the most commonly used distributed file system in Hadoop platform, can store not only single files at TB level, but also tens of millions of files in a file system.

2.3 Construction of Parallel Database.

Shared decision tree algorithm model is still a tree structure, and its construction idea is based on the traditional theoretical basis, which is still an iterative process. Through the analysis of parallelism, the following conclusions can be drawn: according to the structure of the tree, the complete decision tree is divided according to the hierarchy, and there is no direct connection between the nodes in

each layer. From the point of view of data set, the nodes in the same layer are independent of each other and do not interfere with each other, and each of them is a separate data set. Therefore, the corresponding splitting of each node will not affect the operation of other nodes at all. Since the construction of shared decision tree is based on the priority traversal of hierarchical breadth, the brothers in the same layer can be synchronously split to achieve parallel attribute nodes. Parallel attributes among nodes: The process of constructing decision tree is the process of splitting nodes, in which an important measure is used, that is, the attribute splitting index of splitting nodes. Through analysis and understanding, no matter what calculation method is used to calculate the split index, the data used in the calculation is only related to the current node attributes, but not to any other node. Many attributes of all target nodes are independent of each other. Therefore, attributes can be used in parallel to compute the splitting index of multiple attributes at the same time.

Parallel sorting of attributes and nodes: Through the previous explanation, both nodes and attributes can achieve synchronous parallel processing. The process of tree construction needs to calculate the information of nodes, compare and select the best nodes. If we compare and select each two results, the performance of the algorithm will be greatly reduced, which requires corresponding sorting to simplify the calculation complexity. In the process of parallel processing, more parallel sorting operations are needed. According to the independence of nodes and attributes, sequencing parallelism can be realized.

According to the analysis of shared decision tree, the shared decision tree is still a tree with all the structural characteristics of the tree, but its purpose is to mine common knowledge in large data for knowledge sharing. The above analysis shows that attribute parallelism, node parallelism and ranking parallelism exist in the construction of decision tree.

In the decision tree stage of splitting nodes, it is necessary to calculate the corresponding index values according to the rules through the relevant data information of attributes between nodes, find the most suitable nodes for splitting, and split them accordingly. In the large data set environment, there are thousands of attributes. To improve the performance of the algorithm, parallel classification index calculation is necessary. According to the analysis of the parallel principle of Map Reduce, to realize the parallelization operation of attribute tables, it is necessary to separate and store the attribute tables to facilitate the parallelization of attributes among nodes.

By understanding the structure of the tree, tree node splitting is a process of hierarchical splitting. It divides the nodes of each layer recursively, and processes them in parallel according to the characteristics of tree node independence. In the process of Map Reduce processing, the attribute table structure is used to parallelize. By designing the node ID of the attribute on the attribute table and using the mapping mechanism of Map, the following distribution operations are obtained: the attribute table with the same node ID tag will be sent to the same Reduce for calculation, the task of processing the attribute table of the same level node will be completed in parallel, and then the distribution of the attribute table will be done in parallel. The results of table processing are mapped to different Reduces and split at the same time.

Big data technology is an emerging technology which is growing up nowadays. It mainly deals with massive data and extracts hidden information resources. Nowadays, big data technology has become popular all over the world, and its expansion momentum is also inappropriate. It has attracted the attention and research of relevant experts. To extract more useful value from data and satisfy people's demand for computing speed and time of server in real life, the research of classification algorithm based on large data is a very valuable research topic. Traditional classification mining methods have become the bottleneck of the development of large data due to the memory limitation of the machine itself and the lack of computing power. It is difficult to adapt to the processing of large data. Using distributed computing cluster is a good way to solve this problem. As the most popular and efficient data processing platform in distributed cluster, Hadoop has the advantages of high scalable storage, high fault tolerant processing and high concurrent execution.

3. Conclusion

With the continuous development of data technology and the wide use of Hadoop and Spark large data platforms, parallel mining algorithms are required to innovate constantly to adapt to the trend of large data. Based on the classification technology proposed in this paper, there is still room for expansion. Using real-life data to test the performance of the algorithm, continuous optimization and improvement, but also make the algorithm more widely used, although the data is authoritative, it is necessary to use larger and more meaningful data. For example, large medical data, it is very suitable to use the shared decision tree model for analysis.

References

[1] Mkrttchian, V. (2015). Modeling Using of Triple H-Avatar Technology in Online Multi-Cloud Platform Lab. In Encyclopedia of Information Science and Technology, Third Edition (pp. 4162-4170). IGI Global.

[2] Wang, Y., Li, J., & Wang, H. H. (2017). Cluster and cloud computing framework for scientific metrology in flow control. Cluster Computing, 1-10.

[3] Yang, S., Bagheri, B., Kao, H. A., & Lee, J. (2015). A unified framework and platform for designing of cloud-based machine health monitoring and manufacturing systems. Journal of Manufacturing Science and Engineering, 137(4), 040914.

[4] Drawert, B., Trogdon, M., Toor, S., Petzold, L., & Hellander, A. (2016). Molns: A cloud platform for interactive, reproducible, and scalable spatial stochastic computational experiments in systems biology using pyurdme. SIAM Journal on Scientific Computing, 38(3), C179-C202.