Research on Agricultural Big Data Analysis System Based on Hadoop

Lihong Wang¹, Ping Liu¹

¹ Department of computer and information engineering, Heihe University, Heihe, Heilongjiang, 164399, China

*Corresponding author’s e-mail: 116560629@qq.com

Abstract. In order to improve intelligent agricultural management, control and decision-making, in view of the practical needs of agricultural big data application, it is necessary to collect and process large amounts of agricultural data. In order to deal with these massive agricultural data, based on Hadoop large data analysis tool, a large agricultural data analysis and processing system is designed and implemented. Through its powerful distributed data processing capability, various agricultural related data can be analyzed efficiently and quickly. Experiments show that the system can effectively complete large-scale data processing and analysis of agriculture, mining useful information from massive data, so as to store and analyze agricultural data reliably and efficiently. The large-scale data system has important practical significance for improving the level of intelligent agriculture development.

1. Introduction
Agricultural big data is the practice and application of big data theory, technology and method in the field of agriculture [1]. At present, the application of agricultural big data is still in its infancy in China. There are few universally available and universal agricultural big data platforms for agriculture-related fields alone. Some agricultural-related enterprises have real-time or online massive data resources, and need to use large data systems to solve practical application management problems[2]. In order to meet the practical needs and popularization of intelligent agricultural big data, and reduce the cost of use, there is an urgent need for agricultural big data analysis system.

Hadoop is an open source data processing framework of Apache Foundation. The core part is composed of two modules: HDFS (Distributed File System) and MapReduce (Parallel Computing Model) [3]. HDFS can efficiently store large-scale data sets. MapReduce divides the work to be processed by applications into several small pieces, and it is easy for developers to implement distributed application programmers. The combination of the two enables users to write distributed programs at the upper level without knowing the details at the bottom. Users can make full use of the advantages of cluster for distributed and high-speed storage and operation. The combination of HDFS and MapReduce makes Hadoop more powerful[4,5].

This paper designs and implements a big data analysis system based on Hadoop, which uses big data technology to extract value and create wisdom from real-time and complex massive data[6]. The analysis system make the management, control, prediction and decision-making of modern agriculture more "intelligent" and improve the speed and quality of the development of intelligent agriculture.
2. Architecture Design of Large Data Analysis System Based on Hadoop

2.1. Basic System Architecture

The large data analysis system based on Hadoop is divided into the following three layers according to the method of logical hierarchy: data storage layer, data processing layer and application layer. The architecture of the large data analysis system is shown in Figure 1.

Data Storage Layer: Using Hadoop Distributed File System HDFS, the data are classified and stored according to the data characteristics and specific business requirements. A large amount of historical data is stored by HIVE management. The data that needs to be queried and responded quickly is handed over to HBASE for storage partition. The sorted data are saved as files and stored on HDFS distributed file system to facilitate data processing.

Data Processing Layer: This layer is the core layer of large data analysis system. In this layer, we mainly use MapReduce programming framework to build distributed processing program, and use Hadoop components to perform corresponding work to complete various needs analysis of large data. Business logic analysis model in data processing layer is used to identify business functions, which are then decomposed into corresponding tasks to operate on HBase. The model base is used to solidify some commonly used models and analysis results. Middleware judges the specific requirements of application layer by corresponding parameters, and decides to send tasks to Hive according to the result of decision.

Application layer: Data from data processing layer are represented by relevant graphs or tables. Through chart users can intuitively study the characteristics and trends of agricultural data, and strengthen the application of agricultural production.

2.2. Hardware Architecture Design of Hadoop Cluster

Hadoop cluster is composed of 5 servers and 10 PCs. Linux operating system is installed on the server cluster, virtual Linux Cont OS system is installed on 10 PCs, and JDK, SSH, Hadoop and HBase are installed on each machine. A server serves as the client of Hadoop cluster, responsible for data Hbase and Hive storage. One server serves as, amenode, the remaining three servers and 10 PCs serve as data nodes of the HAdoop cluster, two of which serve as middleware servers.

2.2.1. An example.

In this example we can see that there are footnotes after each author name and only 5 addresses; the 6th footnote might say, for example, ‘Author to whom any correspondence should be addressed.’ In addition, acknowledgment of grants or funding, temporary addresses etc might also be indicated by footnotes.
3. Implementation of Large Data Analysis System Based on Hadoop

3.1. Realization of Data Statistics
Large data analysis system needs to analyze and statistics big agricultural data, and the application of MapReduce algorithm framework requires users to write their own programs, which is difficult for employees who use the analysis system to achieve. Hive provides the database language HiveQL, which can transform the SQL-like language into MapReduce tasks performed in Hadoop. Make database operators simple and powerful. The comparison between Hive and general relational databases is shown in Table 1. As can be seen from Table 1, Hive can support large-scale data due to the use of Map Reduce for parallel computing, and has considerable advantages in the scale and scalability of data processed.

| Query language     | HiveQL | SQL     |
|--------------------|--------|---------|
| Data Storage       | HDFS   | Original device or local data storage system |
| data format        | User definition | System decision |
| Data updating      | Unsupported | Support |
| Execute            | MapReduce | Executor |
| Extensibility      | High    | Low     |
| Scale of processing data | Large | Small |
| Hardware configuration | Commonly | High |

3.2. Implementation of Data Storage
In order to improve the storage efficiency and reduce the number of hard disk accesses, first put the data into memory, write to local files when a certain number of data is reached, and upload the data to HDFS through the API provided by Hadoop. The design of HDFS is that each file occupies the whole data block with the highest storage efficiency. In order to improve the storage efficiency and reduce the metadata of Name Node, the local file is controlled at 64M size and uploaded to HDFS.

3.3. Implementation of Data Query
It is like looking for needles in a haystack to quickly locate several or dozens of eligible data in massive data. The application of HBase as a distributed database in large data analysis system can achieve high-speed writing and reading. HBase tables are composed of rows and columns. When querying, they are searched by row keys, so row key planning is particularly important. Data in a region can only be stored on one host. In order to solve the contradiction between reading and writing, adding a hash value before the row key can make the data write to different regions, which can give full play to the advantages of distributed system.

3.4. Data Processing Implementation
Hadoop processes data through MapReduce. The system divides the data into several data blocks. Map nodes will return an intermediate result set after analyzing the data blocks. Some strategies are used to partition the intermediate results to ensure that the relevant data is transmitted to Reduce nodes. Reduce node may process data from multiple Map nodes. In order to reduce data communication overhead, the intermediate results will be merged before entering Reduce node. Reduce node summarizes the received data. Map nodes and Reduce nodes may run in parallel in data processing, even if not at the same time in the same system. Therefore MapReduce provides a simple and elegant solution for data processing in parallel systems.
3.5. Analysis and Implementation of Complex Data Model

In addition to most applications can be satisfied by statistical query, but there is still an important part of the task that needs complex data modeling for analysis. The algorithm steps are as follows:

1. Data extraction.
2. Determine whether the extracted data is in Hadoop, if it is in Hadoop, extract the existing data, if not, import the external data.
3. Data processing and complex algorithm selection.
4. Whether the algorithm is included in Mahout or has been imported, if we turn to step 5, otherwise the imported algorithm will turn to step 5.
5. Setting algorithm parameters.
6. Iterate the algorithm and determine whether the iteration is completed. If the final result is output, otherwise continue step 6.

4. Analysis of experimental results

4.1. Hardware Environment of Platform

The program runs on the Hadoop parallel computing platform, which consists of 22 nodes, including one client, one Namenode and 20 Datanodes. Two servers and 10 Namnodes are configured as eight-core CPU, 64G memory, 400G hard disk, and Gigabit Ethernet network. The other 10 PCs are equipped with dual-core CPU, 8G memory, 100Mbit Ethernet network and 300G hard disk.

4.2. Benchmarking of System Performance

The performance of HDFS is not obvious when the data volume is small, because the control and task scheduling between NameNode and Datanode occupy part of the resources. When the amount of file data is large, the network communication loss can be neglected, and the performance advantage of HDFS can be fully reflected.

The running file size is 100 M, but the number of files increases from 1 to 20. The total running time and average running time are shown in Figure 2.

![Figure 2](image)

Figure 2: The impact of changes in the number of files on runtime

According to the comparison chart, the increase of the number of files leads to the increase of the total running time, but the average processing time shows a downward trend.

Read and write 20 files. The size of a single file increases from 10M to 500M. Figure 3 shows the total run time just and the average run time are shown in Figure 3.

As can be seen from the figure 3, with the increase of file size, the overall time tends to increase, but the average processing time of response (the time for writing 1M files) has been declining.
5. Conclusion
For the characteristics of huge volume of agricultural data, a large data analysis system based on Hadoop is designed and implemented. HDFS is used to design and implement distributed data storage, Hive component is used to complete the statistical task of large data analysis, and HBase distributed database is used to realize high-speed writing and reading of files. The files are distributed to each node reasonably, and the files are backed up at three nodes to ensure the security of the system. Data parallel storage and processing is realized based on MapReduce model of Hadoop. With the increase of data volume, the advantage of cluster data processing becomes more and more obvious. Therefore, Hadoop-based data processing of Russian trade is very safe and effective.

Acknowledgments
Innovation and Entrepreneurship Training Program for College Students in Heilongjiang Province (201813744034)

References
[1] Wu X, Zhu X, Wu G Q. (2014) Data mining with big data. Knowledge and Data Engineering, 26: 97-107.
[2] Allam M, Q W Jia.(2019) Research on Key Technologies of E-commerce Big Data Analysis Platform. Earth and Environmental Science, 5:41-45.
[3] Y Z Guo. (2019)An Improved Parallelization of K-means Algorithm based on HADOOP. Journal of Physics: Conference Series, 4:124-130.
[4] F Y Liang, X S Liu. (2016) Activity of flow patterns in aneurysms on the anterior communicating artery to anatomic variations of the cerebral arterial network. Journal of Biomechanics, 15:340-345.
[5] Allam M, X H Ning. (2019) Research on Online Monitoring Data Storage of Intelligent Substation Based on Hadoop. Earth and Environmental Science, 2:78-82.
[6] Karim D. (2019)Large-scale e-learning recommender system based on Spark and Hadoop. Journal of Big Data, 6:10-23.
[7] M Alsmirat, Y Jalarweh, M Al-Ayyoub. (2018) Speeding DBLP querying using hadoop and spark[J]. Materials Science and Engineering, 1:232-238.
[8] Z J Sun, G H Wang, L B Li, C C Li. (2018) The Study on Bridge Damage Identification Based on Vibration Mode Analysis and Hadoop Platform[J]. Earth and Environmental Science, 1:154-158.