Applications and Implementation of Decomposition Storage Model (DSM) in Paas of Agricultural

Shuwen Jiang, Tian'en Chen*, Jing Dong, and Cong Wang

Department of Information Engineering, NERCITA
Beijing, 100097, China
jiangsw@nercita.org.cn

Abstract. With involvement and the popularization of the Internet of things and cloud computing technology in the modern information agriculture, RDBMS has been difficult to resolve perception storage and analysis of mass data in Internet of things.big data storage and computing are a hotspot and difficulty of research on agriculture cloud computing in recent years. This article is based on agricultural cloud Paas platform, rural areas and farmers services to designe and implemente a distributed storage of large-scale data of perception using the DSM and distributed file system. While providing services of vast agricultural data of perception storage and analysis through agriculture Paas platform, while implementing agricultural cloud computing. DSM is designed based on Hbase as a mass of NoSql database which provides real-time efficient reading and writing, high scalability and high availability. DSM is deploied on the distributed file system of Hadoop which is based on HDFS. While MapReduce of hadoop can provide high-speed large file analysis and processing. Experiments indicate that the DSM technology based on application of agricultural Paas can meet requirement of the perceptive big data, strong scalability and so on.

Keywords: DSM,cloud computing, HDFS, hadoop, Hbase, Paas.

1 Introduction

With the industrial upgrading of agricultural informationization in China, the Internet of things technology and cloud computing is also more and more been used in agricultural products. Big data problem have been followed in the Internet of things technology[1] and cloud computing[2]. Because of the regional distribution is more and more widely, characteristics of wireless sensor network scale is more and more big[3],Sensors' data as the basis most important data of Internet of things technology appeared explosive growth. agricultural big data problem directly appear in front of us. For example, if the sensors' frequency is 5 s, then a sensor got data quantity is 17280 one day, 1000 sensors got data which are nearly 20 million for a day, the amount of data that there are 6.3 billion reasons for a year. Now these big data have

* Corresponding author.

D. Li and Y. Chen (Eds.): CCTA 2013, Part II, IFIP AICT 420, pp. 34–41, 2014.
© IFIP International Federation for Information Processing 2014

saved based on relation or object model database and a large number of small files.But using a relational database and log files storage cost and conventional tool to analyze these data have been faced with resources and operational problem in whole system.

recent years , big data problem is emerging continuously,cloud computing technology also had further development on the analysis of big data. Google first distributed HDFS and storage model. Have developed out of the many open source framework based on Google.which represented by apache Hadoop of distributed computing technology and HBase's DSM. Current agricultural sensors' data type is single, semi-structured and oriented to single data record, therefore DSM technology and distributed computing is suitable for agricultural data processing mode. According to agricultural sensory data are extensively distributed and big, single structure and so on, which used Hadoop to agricultural Paas cloud platform, through DSM technology and HBase open-source database construct storage of Paas. Agricultural cloud Paas platform would put sensors'data storage as a service of platform and opening interface to the outside. the users through the interface to use big data storage services while agricultural Paas cloud also can provide big data analytics service at the same time.

2 Overall Design of DSM on the Paas

2.1 DSM and Hbase

The article "A decomposition storage model" proposed the DSM (decomposition storage model) detail concept at SIGMOD conference in 1985, and the Sybase had DSM Sybase IQ database system 2004 years or so which is mainly used for on-line analysis, query intensive applications such as data mining. DSM compared with the NSM (N - ary storage model), the main difference is that the DSM will all records in the same field and NSM is aggregation of all the fields in each record[4].

The HBase is an open source implementation of DSM. HBase is distributed and the columns of the storage system which provides real-time, speaking, reading and writing and random access to big data sets. HBase table automatically cut into different region, each area contains a list of all the lines of a subset. HBase is composed of a master node server coordination of one or more area (region server). HBase implementation depends on the Zookeeper to coordinate, zookeeper select a node as the Master, the rest of the nodes in the region server. HBase table consists of rows and columns[5], by default, HBase automatically assigns the timestamp when inserted in the cell.content of tables' cells is a byte array which is not explained. Each row of the column are grouped to form columns as column families, all the columns of the cluster members have the same prefix, columns in the group by qualifier.as a result, each column means for the column family: the qualifier.

Agricultural sensory data characteristics is a single structure, large amount of data. Use relational database to build storge in the early stages what did not consider the characteristic of large scale and distributed of sensors'data.and relational database in the bottleneck with the development of the Internet of things technology. general

solution is done by copying or partition distributed storage, but the installation and maintenance cost is very high. Use of distributed storage technology, the expansion characteristics can dynamically add or delete storage nodes without changing the existing data storage way, distribution of the data on the server cluster cleverly. one of the important advantage that agricultural Sensory data used stored in columns is that the entire database is automatically indexed via selection rules are defined by column in the query.when query by only a few fields that greatly reduce the read of big data via storing gathering data of each field for columns storage.while it is more easily to design better compression/decompression algorithm for this storage of gathered.

2.2 Overall Structure of DSM on the Paas

Design a sense data storage service on agriculture Paas platform according to sensors' data type is single, large amount of data.architecture is shown in figure 1. It contains three layers:

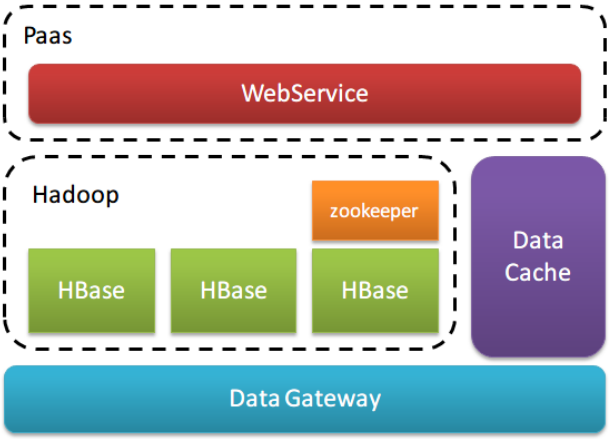

Fig. 1. Structure of DSM on the Paas

(1) the data gateway and the cache: the sensors'data which were got form the Internet of things access to distributed storage cluster via the gateway.the gateway classify the sensors'data accroding to different times and different regions of the data.the gateway also is called data access entrance.If all the area sensors'data are collected to the cluster database at a time, because of the characteristic of the Hadoop cluster to handle large files[6] that will query real time data and analysis take a long time. therefore sensors'data will be collected and stored in the cache from different regions, the cache data is used by the relational database, depending on the time to cache a particular interval data, has been used to carry out real-time query and analysis.

(2)Distributed storage cluster : distributed storage of the sensory data, the hadoop cluster obtain real-time sensors' data from the data gateway and the cache, and have a

persistence. the distributed cluster also manage these metadata and distribution with cluster scheduling.
(3) the WebService service: this mainly provided sensors' data storge service, real-time query and big data analytics services for the agricultural Paas platform. Developers can also use the the sensors'data services which Paas platform provides to customize special application of agricultural data storage and analysis.

3 Application of DSM in the of Paas

3.1 Construct of Hbase on Paas

HBase is a distributed storage application which built on a Hadoop cluster.So first to set up a Hadoop cluster and configure Hadoop cluster. Node number of hadoop's DataNode is 5, the NameNode is 1, Jobtracker node is 1 and a Master node.Each node's CPU is the Intel i3 and 2 Gb ram while have a 500 gb hard drive and the operating system for Ubuntu 10. Use Hadoop version is cloudare open-source, built-in HBase version corresponding to a zookeeper. Set the data block is 64 MB, replications is 3. This Hadoop cluster as a Paas platform is distributed structures for computing platform, data storage. The construction of the Hbase specific steps are as follows:

(1) deployment of zookeeper.zookeeper as Hbase's management scheduling, which is deployed separately on a node.It is used to control the Hbase distributed storage cluster. Zookeeper determine to monitor specific information of Hbase through extracting configuration files.

(2) deployment of Hbase. First to config Hbase on Hadoop NameNode, Hbase is introducesd to the Hadoop cluster; Second to config the associated files of zookeeper and Hbase on Hadoop cluster; Finally, deployment of Hbase in the DataNode and start the Hbase cluster to realize the construction of distributed storage[7].

3.2 Structure Designe of DSM

A Storage architecture is designed by this paper is two structure tables, It respectively contain the sensor group table and sensors'data tables. Sensor group table is mainly used to store sensor group information. Sensorgroupid as a row with the info as a column family.Info provide key/value of key-value pairs to define sensor group information such as name, address, description for the column.Which is including (info: gid, info: IP, info: port, info: region servers, info: avaCapacity, info: the location). Sensors'data table mainly storage sensors' data of sensor group.data table will be sensorgroupid and reverse timestamp as a row, sensordata as a column family and column family contains (sensordata: tem, sensordata: co2, sensordata: soiltem, sensordata: hum, sensordata: soilhum, sensordata: sun)[8].

Efficient reading is the key design of DSM for big data of Sensors'data. For sensor group table, using sensor group id as a row, because it is usually queried as a keyword in the entire sensor group.while the sensory data table row key used a combination of sensor group id and reverse time stamp. so that we can observe real-time data

according to a sensor group data by time stamp. The latest sensory data will be put in front of the row. Table design is illustrated below as Table 1.

Table 1. Design of sensordata table

RowKey	timestamp	timeColumnFamliy : sensordata			
sensorGroupid	time1	sensordata:tem	sensordata:hum	sensordata:sun	sensordata:co2
	time2	temValue1	humValue1	sunValue1	co2Value1
	time3	temValue2	humValue2	sunValue2	co2Value2

3.3 Sensors'data Storage Service

DSM applications of Paas in agricultural is mainly embodied in as a service provided to developers or users.the user encapsulate specific applications by calling service. And sensors' data storage as a service use a three layer scheduling method. First having a persistence at the bottom of the infrastructure.Using Hbase to sensors'data storage.Second processing a the OO storage operation on the Paas layer. The operation of each table in the Hbase needs to acquire the objects of HTable, then use the put and get methods to complete the data operation of insertting and reading. Using HBaseAdmin object to complete the operation of creating and deleting table. Finally At the application level, store operation is encapsulated on the Paas and form the service interface while releasing a version of the WebService.Developers and users can invoke the WebService of storage service to storage sensors'data via the Paas. storage service on Paas as shown in the figure2 below.

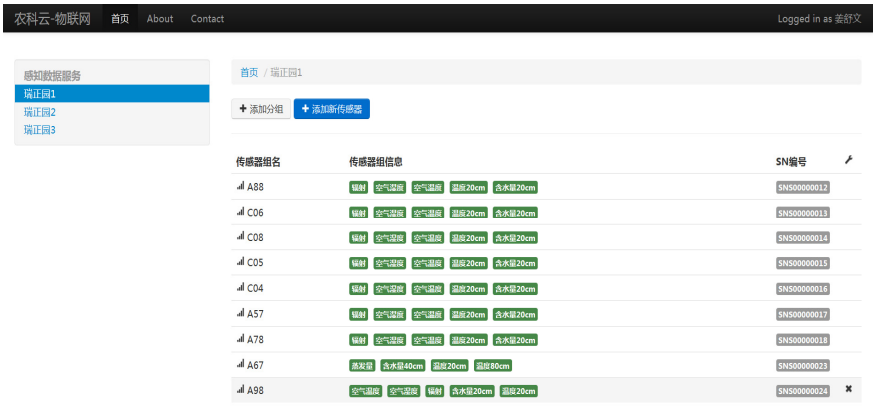

Fig. 2. Storage service on Paas

4 Experimental Results

To verify efficiency of storage and query of HBase and analysis ability of large data, comparing Hadoop cluster which used 20 nodes with single computer in operation of GET and SCAN data's time from the 100 sensor groups for a month. Figure 3 shows the time relationship of GET data between single machine's sensor nodes and cluster.

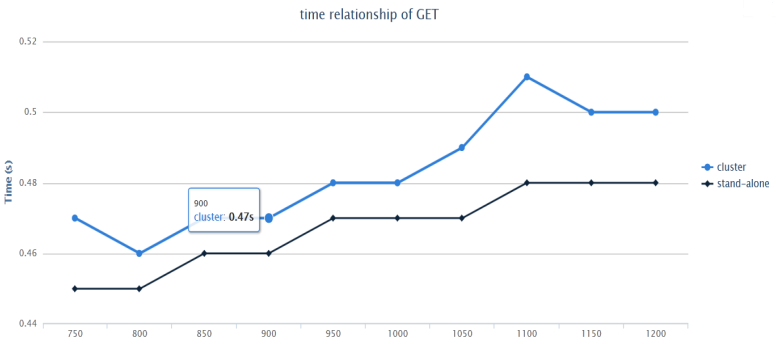

Fig. 3. Time relationship of GET data between stand-alone and cluster

Can be Seen from the Diagram. The gap of standalone and cluster is not big when the amount of data is small. The amount of data increases with the passage of time, the advantage of cluster began to highlight.The growth of reading time is lower than that of single. HBase cluster had an visible obvious advantage for storing big data. Figure 4 shows the time relationship of SCAN data between single machine's sensor nodes and cluster.Can be seen from Fig.4. When cluster and stand-alone execute operation of

SCAN at the same time, efficiency of cluster is lower than stand-alone's. This shows that HBase is not fit to do a full table operation and is suitable for random access.

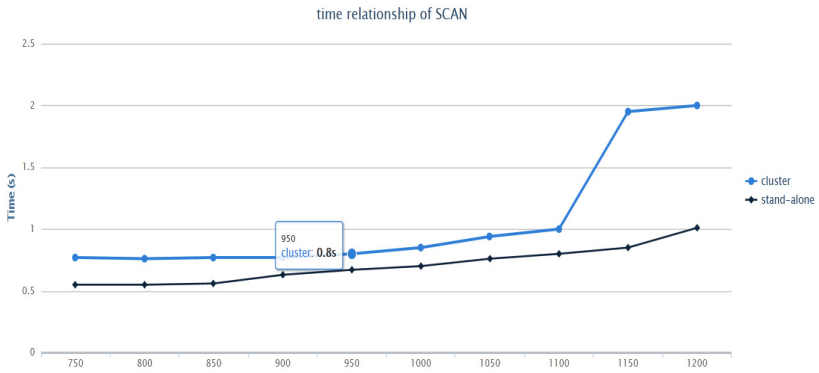

Fig. 4. Time relationship of SCAN data between stand-alone and cluster

For example, analysising nearly the highest temperature for a month. Hbase cluster and single machine carries on a comparison while analyzing by MapReduce in Hbase and analyzing by traditional method in stand-alone.As shown in the figure5 below, the gap of single and cluster is not big when the data volume is less than 1G. Cluster's advantage began to highlight with data volume reaches G level. Analyzing time significantly reduced.HBase cluster got an obvious advantage in analyzing big data .

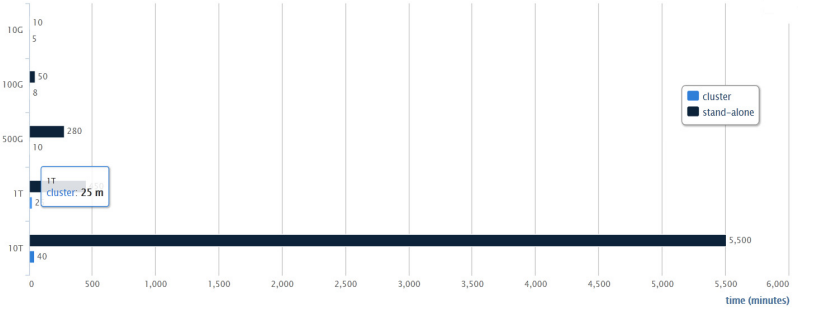

Fig. 5. Time relationship of analyzing between stand-alone and cluster

5 Conclusions

This paper analyzed the problem of sensors'data storage and analysis under the background of big.So putting forward a distributed storage based on DSM architecture and combined with agriculture Paas platform to provide a service. Distributed cluster using Hadoop and Hbase cluster database while putting real-time sensors' data fast persist in distributed file system through reasonable design DSM.

Experiments show that using DSM in the big data persistence has high storage efficiency[9]and having a speed improvements by using computing of distributed cluster.

Acknowledgment. This work was supported in part by National science and technology support project(2013BAD15B05). It also has been done under the help of the team of the Information engineering.

References

1. Conti, J.P.: The internet of things. Communications Engineer 4(6), 20–25 (2006)
2. Yin, G.-M., Liu, Y.-J., Guo, G.-X.: Cloud Computation Applied in the Distributed
3. Yick, J., Mukherjee, B., Ghosal, D.: Wireless sensor networksurvey. Computer Networks 52(12), 2292–2330.6 (2008)
4. Yan, Q.-L., Sun, L., Wang, M., Le, J.-J., Liu, G.-H.: Heuristic Mechanism for Query Optimization in Column-Store Data Warehouse.Chinese Journal of Computers 10 (2011)
5. Chang, F., Dean, J., Ghemawat, S., et al.: BigTablel:A distributedstorage system for structured data. ACM Trans. on Computer Systems 26(2), 1–26 (2008)
6. Hadoop, `http://hadoop.apache.org/`
7. He, Y., Tang, Y., Lin, Y.: Analysis of Construction Profiling Based on Multi-tenant Architecture Paas. Digital Communication, 3 (2012)
8. Zhou, L.-Z., Chen, Q.-K.: HBase-Based Storage System for Wireless Sensor Information of Agriculture. Computer Systems & Applications, 8 (2012)
9. Wang, S., Wang, H.-J., Qin, X.-P., Zhou, X.: Architecting Big Data:Challenges,Studies and Forecasts. Chinese Journal of Computers, 10 (2010)