Applied research on data mining platform for weather forecast based on cloud storage

Haiyan Song[1], Leixiao Li[2*] and Yuhong Fan[3*]

[1] *Department of Software Engineering t, Inner Mongolia Electronic Information Vocational Technical College, China*

[2] *Department of Computer Science of Information Engineering College, Inner Mongolia University of Technology, China*

[3] *People's Bank of China, Inner Mongolia, China*

Received 1 October 2014, www.cmnt.lv

Abstract

This paper analyses the weather data mining, cloud storage and Hadoop framework. A cloud storage platform model of weather data mining is constructed based on Hadoop framework. With this platform two major accomplishments are made: 1)a data model is designed for weather data Mining; 2)a Weather Forecast System is set up through experiment design, data set gathering and prediction algorithm compute. The experimental results showed that this platform is expandable, maintainable, and manages the massive meteorological data with high efficiency

Keywords: (Mass Data Cloud Storage Hadoop Platform Model Weather Forecast)

1 Introduction

With the accelerated industrialization in the world, climate problems such as global warming are becoming more and more serious and the emerging abnormal weather has already caused great economic losses to human society. Meteorology attracts more and more attention. Meteorological data is mainly from the ground meteorological observation stations and air observation stations. Because there are many meteorological stations nowadays, the speed of daily data observations, acquisition and processing is increasing exponentially [1]. However, the storage cost of the massive meteorological data continues to increase. Therefore the effective management of data becomes more and more important,

Which requires that data storage and calculation can reasonably take advantage of inexpensive clusters storage and computing environment to lower the cost effectively manage the existing data and more importantly, could accommodate the increasingly detected meteorological data continuously?

The emergence and development of cloud computing technology provides technical support for large-capacity storage and management of meteorological data. Cloud storage is a dynamic and adjustable storage solution based on Internet. Users could access storage targets through websites under general protocols and application programmer's interfaces. This technology is more beneficial

To the end-user as it can easily increase storage

Capacity and has no requirements to purchase, install and manage any storage infrastructure. This technology is transparent to the end-user in terms of storing and managing large volumes of meteorological data and extending the storage capacity easily [2-4].

In order to store, manage and use these massive meteorological data thoroughly and effectively, this paper used Hadoop as programming framework of weather forecast data mining platform based on cloud storage. Hadoop is an OSS (open source software) under Apache, which makes application program writing and running of massive data easier. The core idea of Hadoop framework is Map/Reduce. Map/Reduce is a programming model used for massive data calculation and a highly efficient task scheduling model which could divide the task into several fine-grained sub-tasks. These sub-tasks can be dispatched between the idle processing nodes. Consequently the faster processing nodes could handle more tasks, thus avoids extending completion time of the entire task due to the slower processing nodes [5]..

2 Cloud storage platform architecture model

Currently, the main problems to store and manage the meteorological observation data are due to the fast growth of data volume and data attribute, fast data response speed upon request, higher security and stability of data and convenient to use and maintain data. Meteorological data storage is raw data and provides data support for subsequent meteorological services, which requires a convenient API interface for the subsequent application program. The system should have a visual interface to facilitate users to read weather data. Additionally, it is easy to maintain the data center equipment. The most convenient is the centralized and unified management of equipment system. Con-

[*] ***Corresponding author's*** e-mail: 342688785@qq.com

sidering the above discussion, cloud storage platform architecture model design of the meteorological data mining is shown in FIGURE 1.

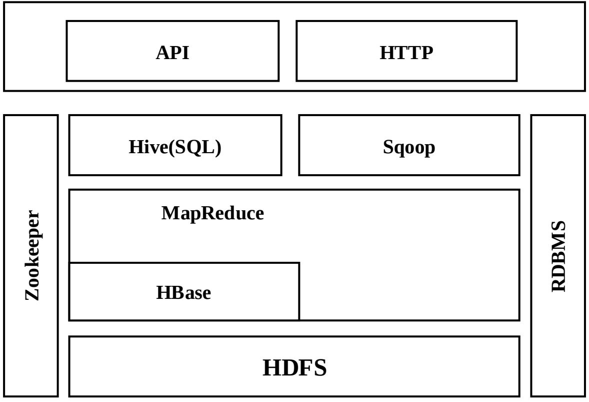

FIGURE 1 Cloud storage platform architecture model

This model is divided into four parts- user access interface layer, metadata storage layer, the entity data storage layer and a relational database [6-8].

1) The entity data storage layer. HDFS of Hadoop is used. HDFS has many advantages such as strong expandability, high reliability and low cost etc. The large files could be stored in general computer clusters at lower cost. But HDFS-the underlying file distribution system of the weather storage platform, is suitable for processing and storing large files. So HBase is used as database and HDFS is the underlying data storage container to store data. HBase does not support the SQL statements and Hive was selected as data warehousing tool. Hive supports Hsql which could make query management operations easier and beneficial to the database operation by API in the late-stage development
2) Metadata storage layer. Metadata storage layer is mainly for storage and management of attributes including HBase database table and Oracle database table. Mapping management of the meteorological database table attribute could be achieved by the management of metadatabase. Metadata table could be stored in HBase.
3) Relational database layer. Oracle relational database could be chosen to store and manage meteorological real-time data and users' data system information of users. Sqoop is used to achieve data transfer and migration. The meteorological data in relational databases is regularly migrated into HBase. The disadvantage of HDFS is the longer response time. For real-time weather data and the smaller amount of data for each point in time. If insert data one by one, the efficiency is lower. Therefore the data with faster response speed should be inserted firstly. And then Sqoop tool could regularly be used for migration. The relational database could be input into the HBase. Finally the storage efficiency could be improved.
4) Access interface layer. HTTP protocol could be provided for WEB service access and API interface could be used for the development and call of other weather-related application businesses.

3 Data model design

Based on investigation and analysis on weather forecast data characteristics and user data demands, database model of the system platform is designed. Database of the system platform includes relational database and distributed database.

3.1 DESIGN OF RELATIONAL DATABASE

Relational database covers three two-dimension relational tables which are user information table, weather data table, and weather forecast data information table. User information table is used to store relevant information of the system users, weather data table can store original weather data from observation stations, and weather forecast data information table will store weather forecast data gained by original data forecast.

3.2 DESIGN OF DISTRIBUTED DATABASE

The date model of distributed database includes weather data model and metadata model, which coveres three tables including Hbase database table of weather, Hive database table of weather, and metadata base table.

1) HBase database table of weather. The row in HBase database table of weather corresponds to the attribute in relational database table of weather, except Timestamp. Timestamp is the time stamp, and its type is 64-digit model. When timestamp is written into Hbase, Hbase will conduct automatic assignment and current system time will be corrected to millisecond [9]. Row in HBase table belongs to one or several row clusters, as shown in TABLE 1.

TABLE 1 Hbase database table of weather

Row key	Time-stamp	Column Family				
		Column1		Column2		Column3
Time+ position	t	time	position	Attribute 1	Attribute 2	Attribute 3

The combination of station and time has a unique value, which will be set as the conditions for query and statistics of weather users.. Therefore, it is used as row key, to provide convenience for query. After integrating Hive and Hbase, the query time is much longer than the time taken by single Hive query or Hbase query. However, by setting attribute as row key for query, the time will be reduced significantly, thus the enhancing the query efficiency.

2) Hive database table of weather .Attribute in Hive database table of weather corresponds to the row in Hbase database table of weather.
3) Metadata base table.Metadata base table mainly manages the attribute for Hbase database table, Hive database table and Oracle database table of

Weather data. It is through operation of metadata base table to manage mapping the relevant weather database table attributes. To map with relational database, field name of weather attribute, type, primary key or not, and non-empty or not is required; to map with Hbase database, field name of weather attribute and the row it belongs to is required; to map with Hive table, field name is required.

4 Experimental scheme design and result analysis

4.1 EXPERIMENTAL SCHEME DESIGN

The weather forecast system is designed and realized bases on the analysis and design of platform and data model; its functional structure can be divided into data acquisition module, result display module, query analysis module, memory module, and data migration[10], as shown in FIGURE2

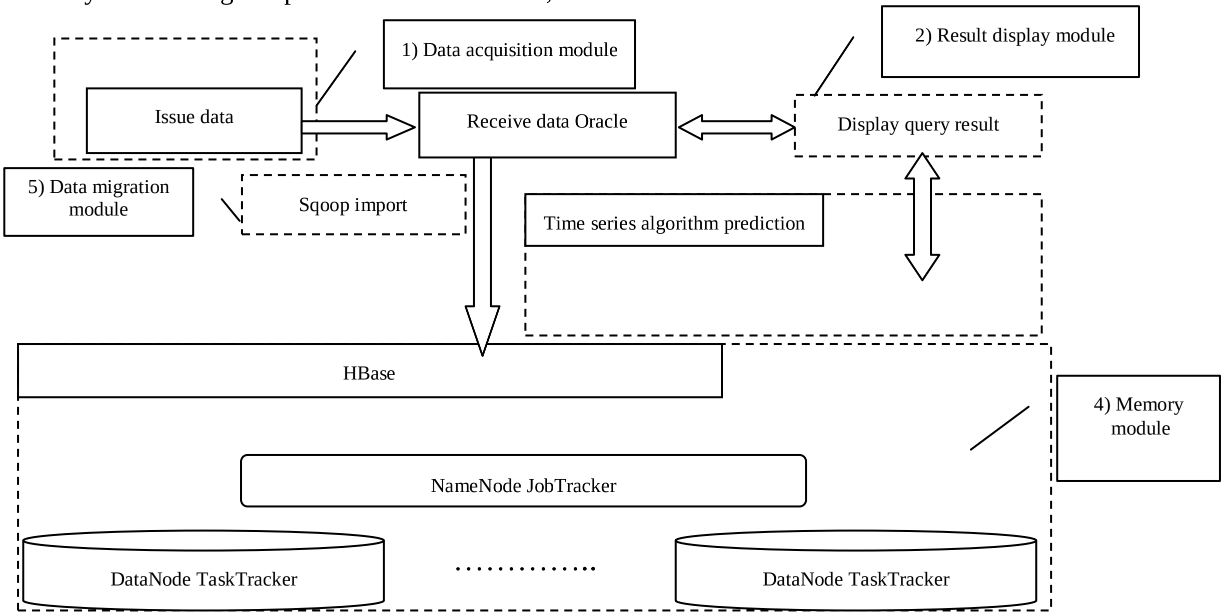

FIGURE 2 Diagram of overall functional structure of weather forecast system

1) Data acquisition module. Visual data dissemination and data acceptance API are provided; weather data can be issued by hand and data can be obtained by switching in data acquisition equipment. The received small data are stored into Oracle database at first; when small data run up to a certain amount, they will be transferred to memory module; the transferred data will be automatically deleted.
2) Memory module. It is responsible for storage of metadata and entity data as well as back-up of data. Hbase is the database of storing metadata and entity data. HDFS is the underlying storage container and its data storage will not be restricted to data types; data of any type are supported. After small data run up to a certain amount, they will be transferred to Hbase of memory module regularly.
3) Query analysis module. Query analysis module includes reading real data and establishing forecasted data. Data warehouse tool of Hive is mainly adopted to read the real data. It supports SQL statement, which is convenient for query. On the contrary, Hbase does not support SQL statement, so developers have to learn language supported by Hbase during the development process, which is quite inconvenient. Hive also provides external data query management API. Hive can automatically translate statement similar to SQL submitted by users into MapReduce. MapReduce is suitable to process big data set; it can obviously increase response speed and return the query analysis result. Establishing forecasted data refers to the function to predict future data, which is realized by ARIMA algorithm. Forecast is made for data in the next 15 days by utilizing data of the past few years. The forecast result will be stored in the forecast table of Oracle database in acquisition module, which can provide convenience for users when querying weather forecast.
4) Result display module. For general users, results returned from query analysis module can be visually displayed in this module; for administrators, it can not only display the query result, but also show the structure of distributed file system. Thus administrators can conduct some management operation for file system structure, and control the database tables.
5) Data migration module. This module uses Sqoop to transfer data in Oracle to HBase, which can be executed by self-timing. Sqoop is a tool to transfer data between Hadoop and relational database; Sqoop imports and exports data via MapReduce, possessing parallelism and fault tolerance. This paper mainly uses Sqoop to import data in Oracle into Hbase database based on HDFS. Table in relational database corresponds to table in Hbase, and row in relational database is corresponds to row in Hbase one by one.

4.2 ESTABLISHMENT OF DATA SET

Experimental data were adopted from ground weather information, including 8 attributes which are station, date, daily average temperature, daily average humidity, daily average vapor pressure, daily atmospheric pressure, daily maximum temperature, and daily minimum temperature. Hbase database table is designed by referring to data model design of Hbase database in 2.2, as shown in TABLE 2.

TABLE 2 Hbase database table

Row key	Time-stamp	Column Family					
		temperature			pressure		humity
attribute	T	AT	MAXT	MINT	AVP	AP	AH

In Table 2, AT denotes avg_temperature (average temperature), MAXT denotes max_temperature (maximum temperature), MINT denotes min_temperature (minimum temperature), AVP denotes avg_water_vapor_pressure (average vapor pressure), AP denotes atmospheric_pressure (atmospheric pressure), and AH denotes avg_humity (average humidity). Attribute is set as Rowkey in the database table; Hbase will carry out automatic assignment for Timestamp when it is written into Hbase. Temperature, pressure and humidity are three row clusters. Each row cluster includes several rows: temperature includes AT, MAXT and MINT; pressure includes AVP and AP; humidity only includes AH.

4.3 ANALYSIS OF EXPERIMENTAL RESULT

Data used in this experiment are weather data collected from 195 ground stations from 1951 to 2011, and 7 data sets with different sizes are intercepted in the experiment: 1.3G, 2.5G, 5G, 8G, 10G, 13G, and 25G.

System response time is selected as the evaluation index. It refers to the overall time

Since client-side issues query commands and data sets being downloaded to the memory model platform until the memory model platform returns the results to client after data manipulation, in the unit of second. Four Experimental schemes are designed in this paper: query for one datum, query for 1 million data, query & statistics for 1 million data, and query & downloading for 1 million data. See Fig. 3, Fig. 4, Fig. 5, and Fig. 6 for the experimental results.

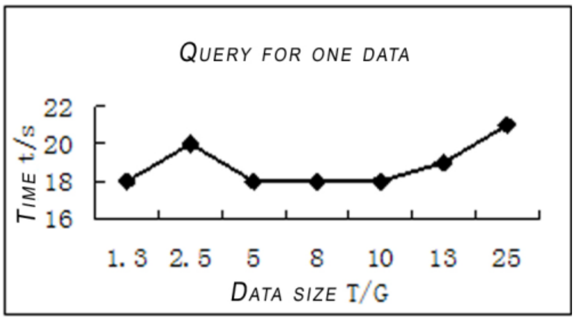

FIGURE 3 Operation result of query for one data

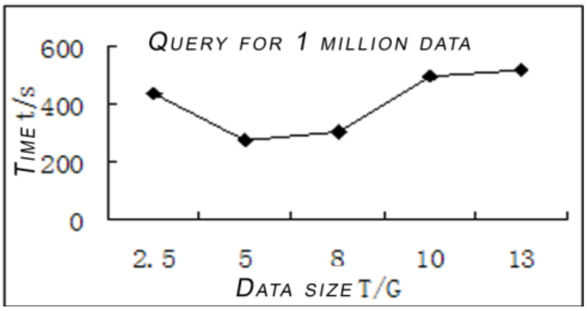

FIGURE 4 Operation result of query for 1 million data

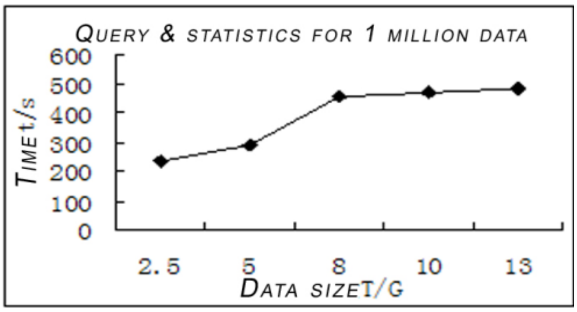

FIGURE 5 Operation results of query & statistics for 1 million data

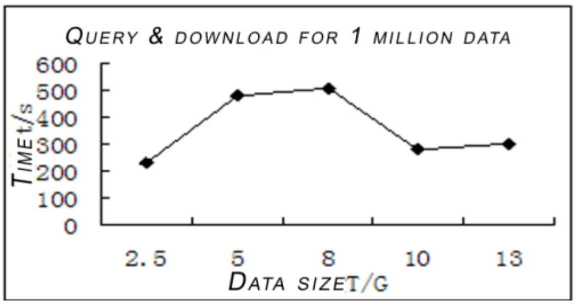

FIGURE 6 Operation result of query & download for 1 million data

According to Fig. 3 and Fig. 4, during data query, when the data set is greater than 5G, system response time decreases obviously. According to Fig. 5, during data query and statistics, when the data set is greater than 8G, the increase range of system response time reduces obviously. As shown in Fig. 6, during query and downloading, when the data set is greater than 8G, system response time decreases obviously. These observations indicate that the system platform possesses great superiority in processing big data sets.

5 Conclusion

Based on the distributed file system HDFS of Hadoop framework, this paper designs and realizes a weather forecast data cloud storage platform in common and cheap computer clusters by combining tools like distributed database HBase, data warehouse management tool Hive, and data migration tool Sqoop between distributed database and relational database. In addition, ARIMA time series prediction algorithm is added into the platform, to realize functions including weather data mass storage, rapid interposition, efficient query and downloading, weather data attribute management, and weather forecast. Thus the shortcomings of low response speed possessed by cloud computing during calculation and storage of big data are overcome. According to analysis of experimental result, this platform is equipped with good expansibility, maintainability, and high-efficient management of massive meteorological data.

References

[1] DOU Yiwen, LU Li, LIU Xulin et al. Meteorological Data Storage and Management System [J]. *Computer Systems & Applications*, 2011, 20 (7): 116-120.

[2] LI Jianguo, YUAN Pingpeng. Study on “Cloud Storage” Scheme Based on Distributed Open Source Management Service [J]. *Computer Applications and Software*, 2010, 28 (10): 208-210.

[3] TAO Changshun. A Cloud Storage Based Mobile Terminal Network Storage Design [J]. *Computer Applications and Software*, 2011, 28 (10): 187-190.

[4] LIU Huanan, WANG Shiqing. Design and Analysis of Data Possession Verification Model in Cloud Storage [J]. *Computer Applications and Software*, 2012, 29 (10): 222-226.

[5] CUI Jie, LI Taoshen, LAN Hongxing. Design and Development of the Mass Data Storage Platform Based on Hadoop [J]. *Journal of Computer Research and Development*, 2012, 49(12): 12-18.

[6] ZHANG Jing, ZHANG Xiaogang. Data Mining Algorithm and Its Engineering Application [M]. Beijing: China Machine Press, 2006: 18-21.

[7] ZHAO Weiying. Time Series Analysis in Meteorology [D]. Yangzhou: master's thesis of Yangzhou University, 2010: 20-23.

[8] ZHOU Ke, WANG Hua, LI Chunhua. Cloud Storage Technology and Its Application [J]. ZTE Communications, 2010, 16 (04): 24-27.

[9] Frank Doelitzscher,Anthony Sulistio,Christoph Reich,Hendrik Kuijs,David Wolf.Private cloud for collaboration and e-Learning services:from IaaS to SaaS.Springer-Verlay New York,Inc.New York, NY,January 23-42,2011.

[10] Sanjay Ghemawat,Howard Gobioff,Shun-Tak Leung.The Google File System[C].Proc.of the 19th ACM Symp.on Operating Systems Principles.2003,P29-43.

Authors	
	Haiyan Song. **She was born in August 1980 and acquired Master's degree in the field of Computer Application Technology from China Inner Mongolia University of Technology in July 2007. Now she is full professor of computer science at Software engineering Department, Inner Mongolia Electronic Information Vocational Technical College. Her current research interests include Data mining; Cloud computing, Software Modelling and Web Information System. In recent years, she has published more than 10 papers about teaching and research in the core journals.**
	Leixiao Li. **He was born in July 1978, and acquired master's degree in the field of Computer Application Technology from China Inner Mongolia University of Technology in July 2007, Since September 2007, he taught in Department of Computer Science of Information Engineering College in Inner Mongolia University of Technology. The main research areas include Cloud Computing, Data Mining, Software Modelling, Analysis and Design.**
	Yuhong Fan. **She was born in Nov. 1987. July 2010, she obtained his bachelor’s degree in computer science and technology from Shanxi Datong University. She studied in College of Information Engineering in Inner Mongolia University of Technology from Sept. 2010 to July 2013. In July 2013, she got the master degree of computer application technology. The main research areas include Data mining, Cloud Computing and Java Application System.**