DESIGN OF CLOUD DATA WAREHOUSE AND ITS APPLICATION IN SMART GRID

H.L. Lv*, F.Y. Wang, A.M. Yan, Y.L. Cheng

*Guodiantong Corporation, State Grid Information & Telecommunication Corporation, Beijing in China, 100070
houlei_lv@sina.com, fengyu_wang@sgcc.com.cn, blue_drop0813@163.com, yanliu_cheng@ sgcc.com.cn
86-10-18201631895

Keywords: Cloud Data Warehouse, OLAP, Data Mining, Smart Grid.

Abstract

With the rapid development of information technology, enterprises have carried out application system construction on a large scale, which generates vast amount of data. Because traditional data warehouse can't meet the demand of processing vast amount of data, the data warehouse based on cloud computing technology was born. The paper puts forward a kind of cloud data warehouse platform and has implemented it. The platform builds on top of Hadoop ecosystem the core of which are Hadoop, HBase, Hive and so on. And the platform also adds ETL, OLAP analysis, Data Mining and BI Report. At the same time, the platform is applied to the field of smart grid in this paper. The paper also uses OLAP to analyse user's power consumption data, and combines data mining technology to solve the problem of analysis of user's power consumption behaviours. Finally, the paper also analyses the defect of cloud data warehouse and the focus of the research direction in the future.

1 Introduction

Hadoop, HBase, Hive and Zookeeper, which are all Apache's projects, develop rapidly. And they make up Hadoop ecosystem. Hadoop ecosystem constitutes the foundation platform of cloud computing, which make it possible that is the efficient processing of TB-level data. Moreover, it has been successfully applied in Yahoo, Facebook, Taobao, and other famous Internet companies. The platform has many advantages, such as economic, reliable, scalable, open source and so on.

Smart Grid Constructions need various intelligent devices, such as smart meters and so on. With the rapid development of Smart Grid Construction, the application of all kinds of intelligent devices become more and more, and the data gathered by grid increase sharply. When the data reach TB-level, traditional data warehouse technology is unable to meet the demand of the grid enterprises' building enterprise data warehouse. It not only has poor scalability, but also is very expensive.

Using scalable, economical and reliable cloud computing technology to build enterprise data warehouse is of great significance. However, the existing Hadoop ecosystem can't meet the needs of building the data warehouse. The paper will discuss the technical details of building data warehouse, and take OLAP analysis of smart grid and analysis of user's power consumption behaviours which is based on data mining technology for an example to introduce the application of cloud data warehouse in smart grid.

2 Building cloud data warehouse

2.1 Whole architecture of cloud data warehouse

Hadoop ecosystem composed of hadoop project and its child project only provides a basal platform applying cloud computing technology, however, building enterprise data warehouse need other kinds of services, such as the interfaces between cloud platform and external systems ,OLAP analysis, knowledge discovery of large scale data, displaying the results of data analysis and so on. According to the exploring in the process of dealing with vast amount of data which is in the smart grid, and summarizing our development efforts, we abstract four modules based on cloud platform. They are ETL module, OLAP module, DM (Data Mining) module and BI Report module. Through integrating cloud computing platform and the four modules, we have proposed the methods of building data warehouse based on cloud computing technology, that is, the construction architecture of cloud data warehouse.

The following is the whole architecture of cloud data warehouse.

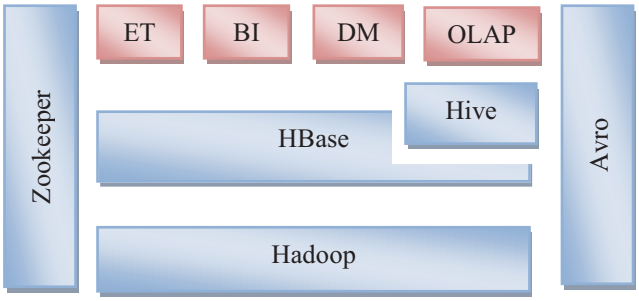

Figure 1: The whole architecture of cloud data warehouse.

In the whole architecture of cloud data warehouse, Hadoop and HBase are underlying foundation, HBase builds on top of Hadoop, Zookeeper and Avro provide coordination service

and data serialization service for the whole platform. ETL module, OLAP module and BI Report module are based on HBase, however, OLAP is based on Hive. And Hive uses its own interface to process the data of HBase.

In figure 1:

(1) ETL: Extract Transform Load is abbreviated to ETL. It offers the interface between various heterogeneous data, such as traditional relational database data, text data, and cloud data warehouse. It supports many kinds of heterogeneous data's being loaded in cloud data warehouse, and it can import the data of cloud data warehouse to relational database. The implementation of ETL module makes full use of Map Reduce distributed parallel computing framework, realize the parallelization of ETL. When ETL module imports the data from relational database to cloud data warehouse, it uses the incremental technique, which greatly improving the efficiency of the data loading.

(2) OLAP: On-Line Analytical Processing is abbreviated to OLAP. Cloud data warehouse's OLAP is realized through Hive, Hive map the column-based table in the HBase to two-dimensional table through its own interface which is between Hive and HBase, and then use HiveQL which is SQL-like language to do OLAP analysis for the data of HBase.

(3) DM: Data Mining is abbreviated to DM. The module provides many kinds of data mining algorithms which are based on Map Reduce computing framework. It offers the user various kinds of knowledge discovery in the cloud data warehouse, such as correlation analysis, predictive analysis, luster analysis, classification analysis and outlier analysis, which provides more valuable knowledge for the users, and supports user to make decision.

(4) BI Report: This module presents the data of cloud data warehouse in the form of visual to the user. The data presented in BI Report come from HBase, the forms of the data's presenting are various. It supports not only traditional list, crosstab, but also graph, bar graph, pie chart and other forms of graphical display.

2.2 Typical data processing flow of cloud data warehouse

Similar to data processing flow of traditional data warehouse, typical data processing flow based on cloud data warehouse includes ETL, data analysis, data displaying. Data analysis includes OLAP analysis and DM analysis. It is in the following.

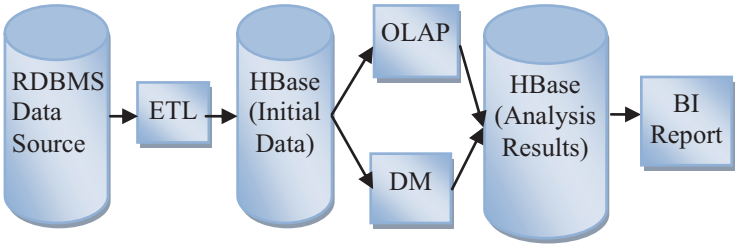

Figure 2: Typical data processing flow chart of cloud data warehouse.

In Figure 2, the data sources of RDBMS are loaded in to HBase through ETL in the cloud data warehouse. Initial data loaded in HBase are analysed and processed in two ways. First, use OLAP module of cloud data warehouse to do multidimensional analysis for the data. Secondly, to use DM module of cloud data warehouse, according to the need of business, select proper data mining algorithms to deal with initial data, and then discovery significant knowledge. The result data sets processed in the above two ways are still stored in HBase, BIReport obtains data directly from HBase, and then presents to the users.

3 Application of cloud data warehouse in smart grid

Smart grid is based on integrated, high-speed bidirectional communication network, through the application of the advanced sensor technology and measuring technology, advanced equipment technology, advanced control method and advanced decision support system technology, achieves the following goals, such as grid reliability, security, economic, efficient, environmentally friendly and safe. Smart grid includes six areas, e.g. power generation, transmission, substation, distribution, consumption and dispatching. With the promotion of smart grid constructions, especially large-scale application of intelligent devices ,such as smart meter, smart home, electric vehicle and so on, make the data in the power grid increase explosively, and the data have achieved PB-level. Facing such vast amounts of data, traditional data warehouse is unable to meet the need of the grid, it that combining cloud data warehouse technology and smart grid becomes an inevitable trend. In the following three chapters, we take OLAP analysis of home user's power consumption data and home user's power consumption behavior analysis for an example to introduce the application of cloud data warehouse technology in the field of smart grid.

3.1 Smart meter and home user electric energy data

Smart meter is the critical device in the field of intelligent power consumption in the process of smart grid construction. We can obtain each user's real-time power consumption data by smart meter. Moreover, if smart meter combines with smart sockets and other devices, we can obtain each real-time data of every user's household electric appliance. Smart meter and smart socket can set different acquisition frequency according to the different needs. Generally speaking, typical home users' power consumption data obtained from smart grid are following: each user's each household electric appliance has 96 records per day, that is, the data acquisition frequency is 15min / time.

At this stage, the home user's power consumption data often are stored in the relational database, in order to load the data from relational database into cloud data warehouse, we need to use ETL module to do parallelization, incremental-oriented data extraction, transformation and loading, and ultimately format the initial data of HBase in cloud data warehouse.

3.2 Home user power consumption OLAP analysis based on cloud data warehouse

Use OLAP module to analyze home user's each household electric appliance loaded in the cloud data warehouse, the analysis is multi-dimensional. The analysis indicator is power consumption the dimensions of which includes time, user of electric and place, electric facilities. The multi-dimensional model of analyzing power consumption is following.

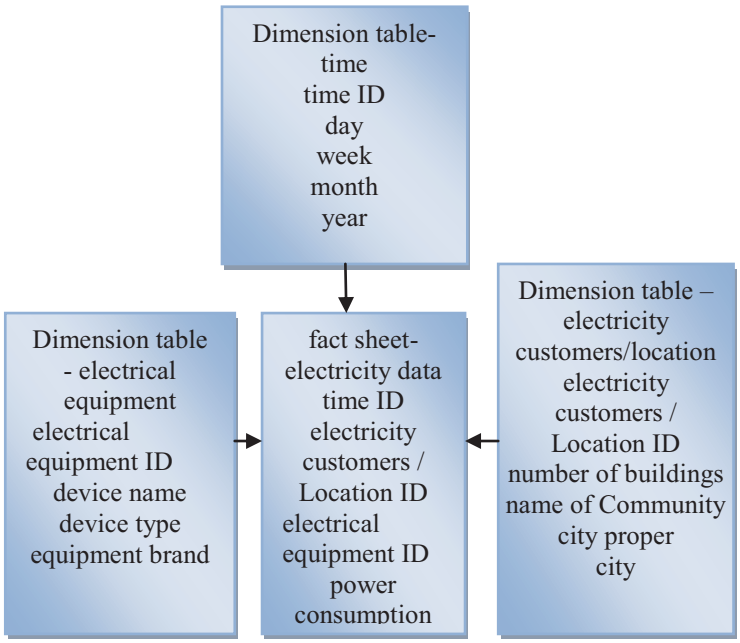

Figure 3: Multidimensional analysis model of home user's power consumption.

Corresponding with the multidimensional model of Figure 3, there are corresponding data tables in HBase. The data tables map the table of HBase to the two-dimensional table, and then it can do multi-dimensional analysis, such as drilling, slicing the home users' power consumption data and so on, through OLAP analysis model of cloud data warehouse. Figure 4 is a multi-dimensional analysis diagram.

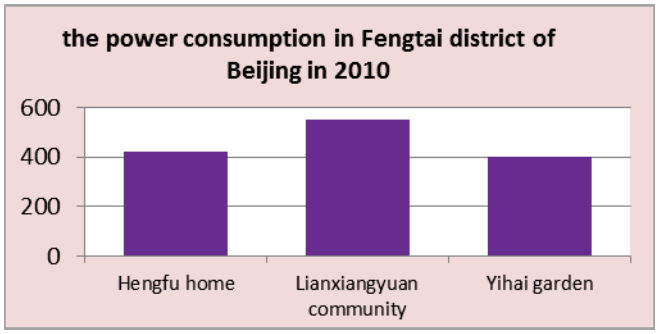

Figure 4: OLAP diagram of home user's power consumption.

3.3 Home user's power consumption behaviour analysis based on cloud data warehouse

The home user's power consumption behaviour analysis is based on each user's power consumption data of per hour, and uses K-MEANS clustering algorithm in the DM module of cloud data warehouse to make clustering analysis. According to the results obtained by K-MEANS clustering algorithm, divide the users into five categories, e.g. A, B, C, D, and E. The power consumption rules of each type of user are different, so we can find the distinctions from twenty four hours' average power consumption law graph.

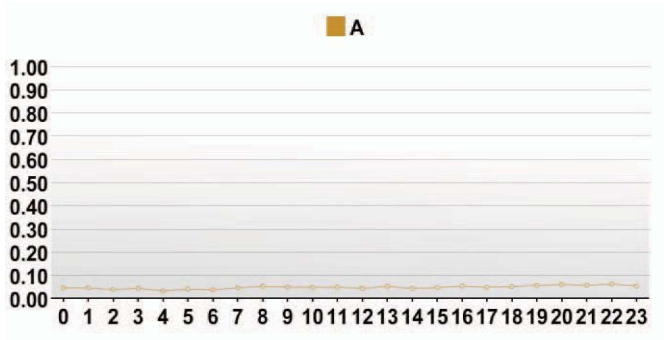

Figure 5: A Class Users' Power Consumption Law Graph.

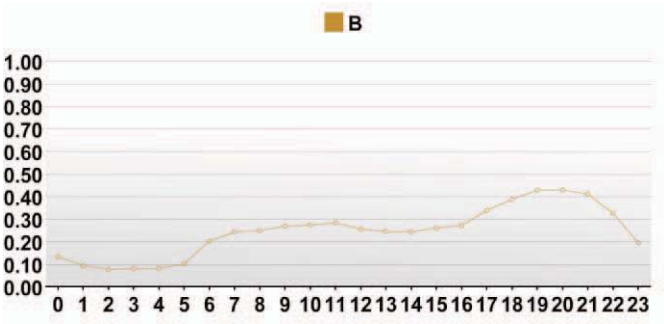

Figure 6: B diagram of home user's power consumption.

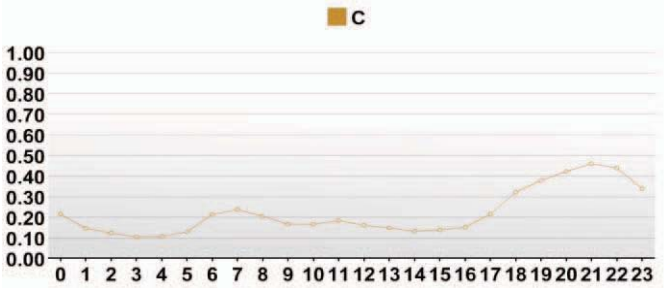

Figure 7: C Class Users' Power Consumption Law Graph.

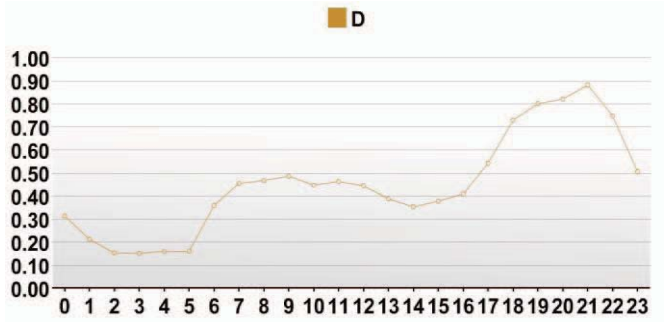

Figure 8: D Class Users' Power Consumption Law Graph.

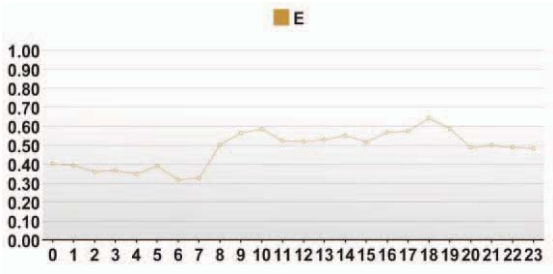

Figure 9: E Class Users' Power Consumption Law Graph.

Figure 5-Figure 9 shows that, A class user's power consumption has been very low in the whole day, so these users may be the vacant house users, and the low power consumption may result from the line loss; B class user's power consumption begins to rise from 6:00, and maintain a certain level during the day, increase at night, but the declining of the power consumption appears earlier. Such users may be older people living in the family; C class user have obvious peak and trough, the power consumption in the day is very low, and the peak of the power consumption in the evening is higher than B, the time of the downward trend appeared later than the class B. Such users may be the office workers living in the household; D class combines B with C, such users may be mixed families composed of older and workers; E class user's power consumption has been very high during the whole day, such users may use the resident houses for commercial purpose.

4 Deficiency of cloud data warehouse

Cloud data warehouse discussed in this paper has been found the following deficiencies in the process of the practical application.

(1) Lack of access control to cloud data warehouse.

The users can access all the data of cloud data warehouse, without any restriction, as long as they can connect to the cloud data warehouse, which is unimaginable to the data warehouse products which are now available in the fact.

(2) The function of cloud data warehouse in daily management is far from perfect.

In addition to various data analysis mode, data warehouse should provide the basic daily management functions, such as setting regular tasks etc. Cloud data warehouse still has many shortcomings in this respect.

(3) Cloud data warehouse lacks unified data presenting service.

Generally speaking, cloud data warehouse is divided into two parts, the background and foreground. The background takes charge of the analysis and processing of the data, and the foreground takes charge of presenting the analysis results. Cloud data warehouse haven't unified data presenting service, the existing BI Report module only obtains data through the access interface provided by HBase, and displays the data. The combination of the module with cloud data warehouse is loose and less efficient.

(4) Cloud data warehouse lacks unified management interface.

Cloud data warehouse's service module, e.g. Hadoop, HBase, Hive etc., perform configuration management through their respective command line tools. The whole cloud data warehouse lacks unified management platform and the entrance.

5 Conclusion

Data warehouse based on cloud computing technology uses Hadoop ecosystem to realize the distributed storage and parallel computing of TB-level vast amount of data. The tool sets which include ETL, OLAP analysis, data mining and BI Report etc., and are provided by cloud data warehouse, have provided various aspects of services, such as data loading , multi-dimensional data analysis, knowledge discovery and data analysis results showing. These services have enriched the functions of cloud data warehouse and make cloud data warehouse practical. Cloud data warehouse can expand its own storage space and computing power by adding ordinary PC nodes, and then meet the need due to the increasing of the amount of data and the performance requirements. Because cloud data warehouse can be expanded by adding low-cost PC nodes, the cost of cloud data warehouse's expanding can be controlled in a reasonable range.

It must be noted, however, there are still many defects in the cloud data warehouse, such as metadata management, data display services, unified management interface and other aspects. The above aspects have vast improvement room. It is the focus of our research in the future.

References

[1] Z. Y. Liu. "Smart Grid Technology", *China Electric Power Press*, 2010.

[2] J. Y. Wu. "The Application Review of Data Mining and Data Warehouse Technology In E-commerce", *Fu jian Education institute journal*, **4**, pp. 126-128, (2010).

[3] J. W. Han. "Data Mining: Concepts and Techniques", *Machine Press*, (2007).

[4] S. Ghemawat, H. Gobioff, S. T. Leung. "The Google File System", *OSDI*, (2003).

[5] Jeffrey Dean, Sanjay Ghemawat. MapReduce: Simplified Data Processing on Large Clusters", *OSDI*, (2004).

[6] F. Chang, J. Dean, S. Ghemawat, et al. "Bigtable: A Distributed Storage System for Structured Data", *OSDI*, (2006).

[7] "The Apache Software Foundation", *http://hadoop. apache. org/*, (2008).

[8] "The Apache Software Foundation", http://hbase. apache. org/, (2011).

[9] "The Apache Software Foundation", *http://hive. apache.org/*, (2010).

[10] "The Apache Software Foundation", *http://mahout. apache.org/*, (2011).

[11] loudera, Inc. *http://archive.cloudera.com/cdh/3/sqoop /SqoopUserGuide.html*, (2011).