

The Construction of Cloud Data Analysis Platform and Its Application in Intelligent Industrial Park

Ning LI*, Pengzhan FAN**, Houlei LV*

Guodiantong Corporation, 28F Ages Wealth World, No1 Hangfeng Road, Fengtai, Beijing, China

ning_li@sgcc.com.cn, pzf@sgcc.com.cn, houlei_lv@sgcc.com.cn

Abstract—The data of smart grid have high frequency of collection, large data size as the characteristics, so traditional technical methods can't meet the demand of the data processing of smart grid. The distributed parallel computing framework the core of which is MapReduce provides the basis for the construction of smart grid cloud data analysis platform. The paper proposes and has actually built a cloud data analysis processing platform used in the smart grid environment. The platform uses MapReduce technology to achieve the data loading of smart grid, multi-dimensional analysis and knowledge discovery based on data mining technology. At the same time, the paper uses the platform to construct an intelligent industrial park, which has solved some problems, such as the storage of massive data of the park, load forecasting and enterprise power consumption behaviour analysis and so on. Finally, the paper also analyzes the defect of cloud data analysis processing platform and the focus of the research direction in the future.

Keywords— Cloud Computing, Distributed Computing, Smart Grid, Data Mining, Intelligent Industrial Park

I. INTRODUCTION

Smart grid is based on integrated, high bidirectional communications network. And it uses the application of advanced sensor and measurement technique, the advanced equipment technique, the advanced control method and the advanced decision making support technique to make grid get the following features which are the reliability, the security, the economic, the efficient, the environmentally friendly and the security. Chinese Smart grid can be divided into six fields: power generation, transmission, transformation, distribution and power grid dispatch. With the promotion of smart grid construction, intelligent industrial park, intelligent community and electric vehicle charging/charging station are established, 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. How to store these data and use them more effectively has become the key of the construction of smart grid. With the increasing of the data, the disadvantages of traditional data storage and analysis processing technique become more and more obvious, such as poor scalability, data processing performance degradation, the high cost and so on.

The cloud computing technique the core of which is distributed storage and distributed parallel computing, can

integrate many cheap ordinary PC into a cloud computing cluster to deal with massive data, and the cloud computing cluster can be extended very easily. The open source project Hadoop of Apache Foundation has implemented distributed file system HDFS (Hadoop Distributed File System) and MapReduce parallel computing model, which provide the basic platform for cloud computing. Hadoop and its child project HBase, Hive, Pig, Zookeeper, Avro make up Hadoop ecosystem, which provide the basis for the distributed parallel technique's becoming practical. China Mobile, Yahoo, Taobao, Facebook and other famous companies all use Hadoop ecosystem to carry out practical application.

Using scalable, economical and reliable cloud computing technology to construct the data analysis platform of smart grid is of great significance. However, the existing Hadoop ecosystem can't meet the demand of constructing the data analysis platform. The paper will discuss the technical details of cloud data analysis platform, and introduce how to use cloud data analysis platform to store massive data, and do load forecasting and enterprise power consumption behavior analysis and so on in the process of construction of intelligent industrial park.

II. THE CONSTRUCTION OF CLOUD DATA ANALYSIS PLATFORM

A. The whole Architecture of Cloud Data Analysis Platform

Hadoop ecosystem which is composed of Hadoop project and its child project only provides a basal platform applying cloud computing technique, however, building enterprise data analysis platform needs 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 cloud data analysis platform based on cloud computing technology, that is, the architecture of cloud data analysis platform.

The following is the whole architecture of cloud data analysis platform.

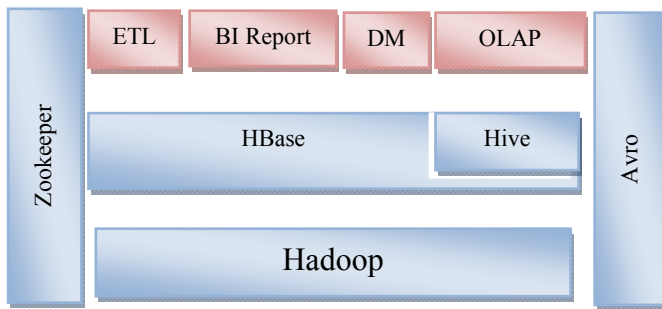


Figure1. the Whole Architecture of Cloud Data Analysis platform

In the whole architecture of cloud data analysis platform, 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:

- 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 analysis platform. It supports many kinds of heterogeneous data's being loaded in cloud data analysis platform, and it can import the data of cloud data analysis platform 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 analysis platform, it uses the incremental technique, which greatly improving the efficiency of the data loading.
- OLAP: On-Line Analytical Processing is abbreviated to OLAP. Cloud data analysis platform'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.
- 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 analysis platform, such as correlation analysis, predictive analysis, cluster analysis, classification analysis and outlier analysis, which provides more valuable knowledge for the users, and supports the users to make decision.
- BI Report: This module presents the data of cloud data analysis platform 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.

B. Typical data Processing Flow of Cloud Data Analysis Platform

Similar to data processing flow of traditional data analysis platform, typical data processing flow based on cloud data analysis platform includes ETL, data analysis, data displaying and other process. And data analysis includes OLAP analysis and DM analysis. It is in the following figure.

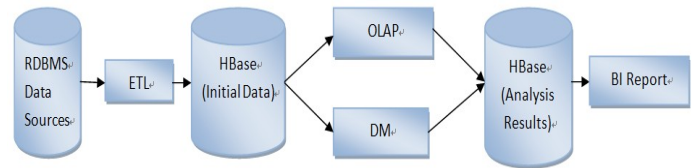


Figure2. Typical Data Processing Flow Chart of Cloud Data Analysis Platform

In figure 2, the data sources of RDBMS are loaded in to HBase through ETL in the cloud data analysis platform. Initial data loaded in HBase are analyzed and processed in two ways. Firstly, use OLAP module of cloud data analysis platform to do multidimensional analysis for the data. Secondly, using DM module of cloud data analysis platform, according to the needs of business, select proper data mining algorithms to deal with initial data, and then discover significant knowledge. The result data sets processed in the above two ways are still stored in HBase, BI Report obtains data directly from HBase, and then presents the analysis results to the users.

III. THE APPLICATION OF CLOUD DATA ANALYSIS PLATFORM IN THE INTELLIGENT INDUSTRIAL PARK

A. the Storage of the Intelligent Industrial Park Power Consumption Data

In the process of the construction of intelligent industrial park, smart meters, temperature sensors and other smart devices finish the collection of the data with PLC (Programmable Logic Controller). And the data are loaded into the cloud data analysis platform through the child module provided by ETL. The child module is used to interact with PLC. Throughout the process, PLC reads the data from smart meters and temperature sensors, and then writes in these data to HBase through the child module which belongs to ETL and is used to interact with PLC. Each PLC can control the data collection of 20~30 intelligent device multiple PLC according to the installed number of its intelligent devices. In the intelligent industrial park, the acquisition frequency of intelligent devices is one time every 5 seconds.

B. the Enterprise Load Forecasting of Intelligent Industrial Park

Load forecasting is of great significance to the grid scheduling and ensuring the safe operation of power. The power consumption of one particular enterprise is relevant to multiple kinds of factors, such as hour, month, workday/weekend, temperature and so on .In order to ensure 24hours power consumption forecasting of enterprise to be more accurate, need to be on the basis of the history load data of the enterprise, take into account the effects of hour, month,

workday/weekend, temperature and other factors on the power consumption, build a reasonable prediction model.

In the process of forecasting the enterprise power consumption of intelligent industrial park, we have implemented locally weighted linear regression algorithm to meet the need of load forecasting, and the above algorithm is implemented according to the combination of history load data and temperature data with the weather forecasts data. The history load data are each hour data of every enterprise collected from the cloud data analysis platform. The temperature data are every hour data collected by temperature sensors. The weather forecasts data is collected from China Meteorological Bureau .LWLR algorithm have following characters:

- The data include five dimensions: corporate number, hour, month, workday/weekend, temperature.
- Every dimension has its own particular distance computing methods. Take the enterprise number dimension for an example. When the numbers are same, the dimension distance is zero, but when the numbers are different, the dimension distance is Infinity. The distance of hour dimension is calculated according to the ring the circumferenceof which is 24.the total distance of two data is the sum of the distance of each dimension.
- Obtain the “local” through KNN algorithm.
- Use Gaussian mixture model to calculate the weight of local data points
- The algorithm executes parallel on cloud data analysis platform, which is efficient. And the platform is easy to expand.

The load forecasting results are stored in HBase, and shown through BI Report. The following diagram is the load forecasting curve of 24 hours of the day of an enterprise.

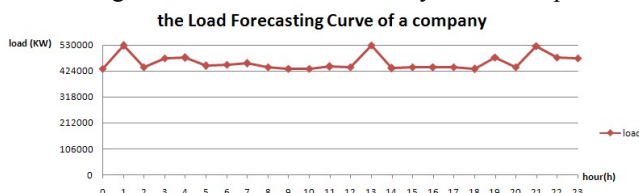


Figure3.the load forecasting curve of a company

C. the Enterprise Power Consumption Behaviour of Intelligent Industrial Park

In our practice, the analysis of the enterprise power consumption behavior in the intelligent industrial park are based on the electricity power data of every hour of each enterprise, use the cluster algorithm K-MEANS of DM module in the cloud data analysis platform to make cluster analysis for the electricity enterprise. The figure 4 is two types of power consumption law tendency chart of the cluster analysis results , horizontal axis represents the hours, vertical axis represents the power consumption. The difference of the two types of electricity enterprise can be seen clearly from the following figures.

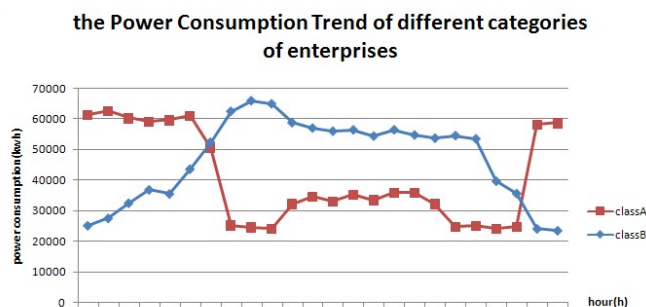


Figure4.the Power Consumption Trend of different categories of enterprise

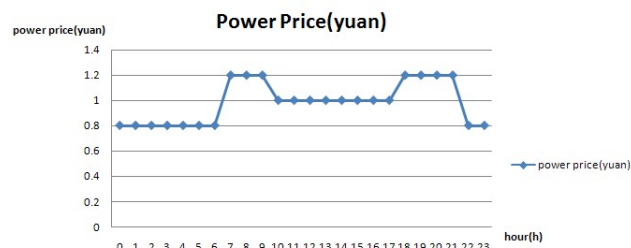


Figure5.the Power Price Trend graph

The figure 4 shows that A, B, two types of enterprises have sensible difference in the power consumption law. A enterprise's power consumption peak appears from 22:00to 6:00in the evening, however, from 07:00to 21:00 in the day is the low ebb of the power consumption; while B is just contrary to A. the peak appears during the day, and the low ebb appears in the evening. If combine with the power price of the day provided by figure 5, A enterprise's power consumption peak appears when the power price is in the lowest point. However, when the power price is growing, A enterprise's power consumption drops off significantly, which is very beneficial to reduce the electricity of A enterprise. On the contrary, when the power price is high, B enterprise's power consumption also is very high, and when the power price is on the decline, the power consumption drops off, which is against B enterprise's cutting down power price. So we advice that B can change working hours, arrange it at the night, because the power price is low in the evening.

IV. THE DEFICIENCY OF CLOUD DATA ANALYSIS PLATFORM

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

- The Lack of access control to cloud data analysis platform. The users can access all the data of cloud data analysis platform without any restriction, as long as they can connect to the cloud data analysis platform, which is unimaginable to the data analysis platform products which are now available in the fact.
- The function of cloud data analysis platform in daily management is far from perfect. In addition to various data analysis mode, data analysis platform should

provide the basic daily management functions, such as setting regular tasks etc. Cloud data analysis platform still has many shortcomings in this respect.

- Cloud data analysis platform lacks unified data presenting service. Generally speaking, cloud data analysis platform 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 analysis platform haven't unified data presenting service, the existing BI Report module only obtains data through the access interface provided by HBase, and display the data. The combination of the module with cloud data analysis platform is loose and less efficient.
- Cloud data analysis platform lacks unified management interface. Cloud data analysis platform's service module, e.g. Hadoop, HBase, Hive etc, perform configuration management through their respective command line tools. The whole cloud data analysis platform lacks unified management platform and the entrance.

V. CONCLUSION

Data analysis platform based on cloud computing technology uses Hadoop ecosystem to realize the distributed storage and distributed Parallel Computing of vast amount of data. The tool sets which include ETL, OLAP analysis, data mining and BI Report etc, and are provided by cloud data analysis platform, 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 analysis platform and make cloud data analysis platform practical.

Cloud data analysis platform 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 analysis platform can be expanded by adding low-cost PC nodes, the cost of cloud data analysis platform's expanding can be controlled in a reasonable range.

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

REFERENCES

- [1] Zhenya Liu. Smart Grid Technology[M], China Electric Power Press, 2010
- [2] Jinyan Wu. The Application Review of Data Mining and Data Warehouse Technology In E-commerce[J], Fu jian Education institute journal
- [3] Jiawei Han. Data Mining: Concepts and Techniques[M], Machine Press, 2007.
- [4] Sanjay Ghemawat, Howard Gobioff, Shun-Tak Leung. The Google File System[J], 2003.
- [5] Jeffrey Dean, Sanjay Ghemawat. MapReduce: Simplified Data Processing on Large Clusters[J], 2004.
- [6] Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach, Mike Burrows, Tushar Chandra, Andrew Fikes, Robert E. Gruber. Bigtable: A Distributed Storage System for Structured Data[J], 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] Cloudera, Inc. <http://archive.cloudera.com/cdh/3/sqoop/SqoopUserGuide.html>, 2011.