Design of Electric Power Data Management System Based on Hadoop

Yongheng Li[1, a], Yongzhi Wang[1,b] and Liang Jin[1,c]

[1]College of Instrumentation & Electrical Engineering, Jilin University, 130061 Changchun, Jilin China

[a]18946628992@163.com, [b]iamwangyongzhi@126.com (corresponding author), [c]1161532932@qq.com

Keywords: Hadoop，Electric power data，HDFS，MapReduce

Abstract. With the development of smart grid, electric power system data surges in a short period of time, and the traditional data storage and processing platform can't meet the requirements. In this paper, big data-related technology will be applied to the electric power data management, using Hadoop as a basic platform. The system stores data by the HDFS distributed file system, through MapReduce framework to achieve data processing parallelism. Object-oriented programming language Java is used to develop the prototype of the system. The system can be applied to power decision support, power system monitoring, power data management and many other aspects. This paper provides a valuable reference for the electric power system data management.

Introduction

In order to ensure the stable operation and timely monitoring management of the electric power system, it is necessary to provide sufficient data support by using a variety of business subsystems. Based on this reason, a large number of monitoring data and historical data which can effectively reflect the operation status of the electric power grid are generated during operation. The electric power system is faced with the following aspects: (1) Large amount of data. The scale of the collected data will exponentially surge to TB or PB level. (2) Diversification of data types. The diversity of data types requires the diversity of storage and processing technologies. (3) High speed. Through the relational database to store, both the write speed and query efficiency will be difficult to meet the needs of the application[1-3].

Facing to the mass, conventional data storage and management methods will encounter great challenges. Based on the analysis of data sources, structure and characteristics of electric power system, combined with large data integration management technology, the system puts forward the electric power data management system based on Hadoop, which make full use of computing resources. It realizes the reliable storage and management of the whole business information of the electric power system with low cost, high reliability, easy to expand and other advantages.

System framework based on Hadoop

Electric power data management system consists of two parts. The first is the integrated management of the electric power data, the complex data environment requires to extract and integrate the data source data. Using non-relational database for storage, the overall relationship will be divided into logical segments from the logic of the overall relationship[4]. The second is the data analysis and processing, using the Apache Hadoop project, which is open source, scalable and distributed application computing architecture[5-7]. Management system consists of four layers: storage layer, computing layer, control layer and application layer, as shown in Figure 1.

Specific description：

(1) Storage layer. At first, the collected data is transferred to the pre-data buffer queue, and put the pre-processed data into the storage layer. The storage layer holds all the data of the electric power system, including real-time production data and static data such as personnel and equipment. Low efficiency of static data stores in SQL Server and other relational database such as the staff and equipment; large and volatile data such as real-time production data is stored in HDFS.

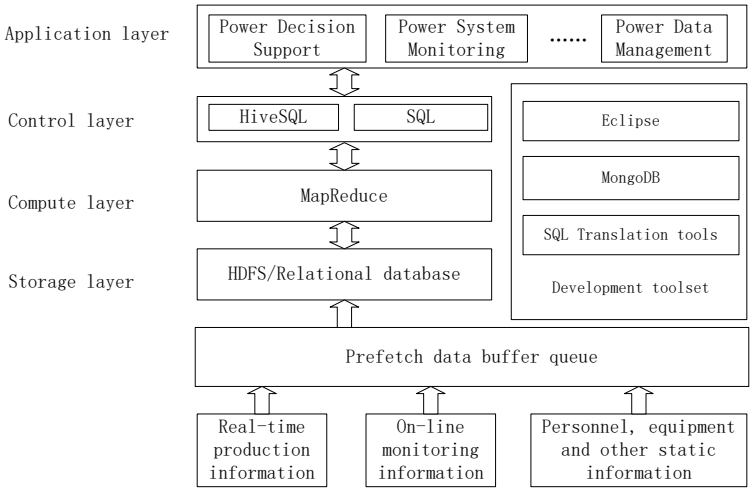

Figure 1 Framework of management system based on Hadoop

(2) Compute layer. The compute layer uses MapReduce, a low-level tool of Hadoop, as a method for computing tasks in data warehouses. Electric power system database computing tasks include historical data query, multi-dimensional analysis, report generation, incremental maintenance and metadata access. The compute layer can decompose any computational tasks into Map and Reduce jobs and dynamically allocates them to different nodes for parallel execution[8].

(3) Control layer. The control layer consists of two query languages, HiveQL and SQL. The database engine handles different requests from the application layer and generates computational tasks to the compute layer. HiveQL is used to parse the query statement of the data warehouse. After the HiveQL analyzing, the MapReduce job is generated and carried out the task by compute layer. The result is generated by Hive's user interface back to the client.

(4) Application layer. The application layer mainly integrates the functional components such as PMS application, auxiliary decision-making and condition monitoring, and realizes the query, analysis and decision of status information.

Application Instance

Functional Model of Electric Power Data Management System

In the electric power system, the classification and fusion design of information data is the key. Therefore, the management system designed in this paper includes four parts, namely electric power generation part, electric power transmission part, electric power consumption part and user management authority part. The overall function of the system is shown in Figure 2.

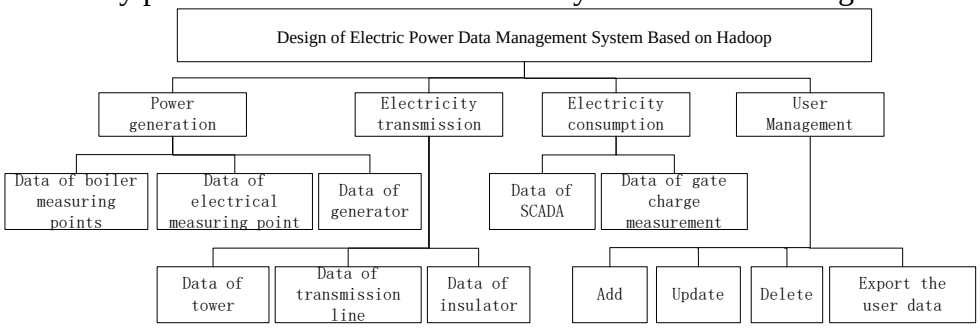

Figure 2 Functional model of electric power data management system

(1) Dynamic timing data: The dynamic data of electric power system refers to the data which can reflect the real-time running status of the electric power system. Dynamic data can be divided into two categories: analog and digital, analog is to monitor the continuous measurement of dynamic data points, the switch is to monitor the discrete points of dynamic data. The characteristics of such data is a very significant growth. With the growth of time required storage space is growing, and gradually beyond the traditional database processing capabilities, and the amount of data to be processed is

huge in the data query. Therefore, the dynamic data is stored in MongoDB through the use of MongoDB database API interface, combined with MapReduce to achieve efficient parallel query.
(2) Static data: Static data in the electric power system includes user data, tower data, acquisition equipment data, line data, insulator data and so on. Because of the need for continuous queries, including MongoDB is not good at the Join operation, so the static data is still stored in the relational database SQL Server.
(3) Unstructured data: As the image resolution of video data is very important, the higher the resolution, the more obvious details, the higher the accuracy of monitoring. In this case, the use of HDFS file system for storage provides a great convenience. Therefore, the focus of this paper is to achieve the storage and access to the cluster in the Hadoop platform based on SOA, through the HDFS and MongoDB API, as well as MapReduce procedures and Java programs.

System Implementation

The Hadoop platform runs on Linux, Using Ubuntu Server 14.04 version of the Linux system, and the platform consists of four virtual machines to build a total of 7 core (1x7).Virtual machine builds on the HP server, with 12 CPU, 96G memory, 2T hard drive. Then the cluster software installation and network and parameter configuration is carried out, the cluster structure shown in Figure 3. The following is an example of generator information management. Figure 4 shows the generator information management interface, which belongs to the power generation module. The basic parameters of generators are static data, which are stored by relational database. However, the real-time production data generated by generators are stored in MongoDB. In this interface, we can add, delete, modify, query and analyze the operation of information for the generator.

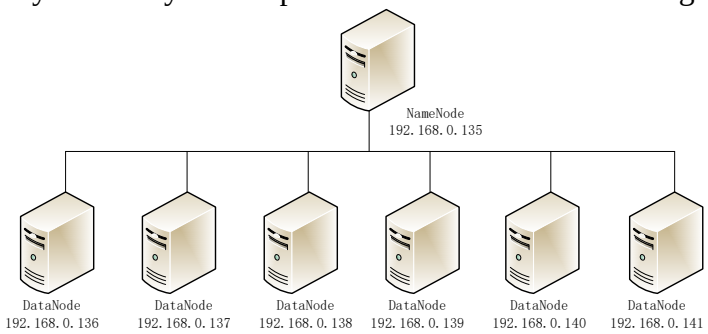

Figure 3 Cluster structure diagram

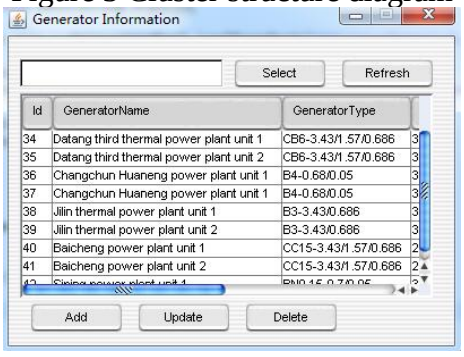

Figure 4 Graphical user interface of generator information

Summary

In order to meet the demand of electric power system data management, this paper develops a management system based on Hadoop for the typical characteristics of electric data, combined with the latest development of big data technology and actual deployment experience. The system is based on the distributed parallel computing framework (Hadoop), using HDFS to store data and reading data based on MapReduce. It makes full use of the advantages of distributed computing, and realizes the storage and analysis of electric power system business information. The system provides the means to dig deeper into the potential value of electric data.

References

[1] Wang, Y., Gao, G., Yang, Y., et al: Technology for the construction of the geoscience spatial data warehouse. Geological Bulletin of China. 27(5):713-718 (2008).

[2] Wang, Y., Zhang, Y., Yi, J., Qu, H., Miu H.: A Robust Probability Classifier Based on the Modified chi^2-Distance. Mathematical Problems In Engineering. 2014:1-11(2014).

[3] Wang, Y., Pan, M., He, W.: Design and Implementation of National Oil-Gas Resource Database Management System Based on ArcGIS and SOA. Journal of Jilin University(Earth Science Edition). 39(5):953-958 (2009).

[4] P. Atzeni, F. Bugiotti, L. Rossi: Uniform access to non-relational database systems: the SOS platform. (International Conference on Advanced Information Systems Engineering, Poland 2012).

[5] V.K. Vavilapalli, A.C. Murthy, C. Douglas, et al: Apache Hadoop YARN: yet another resource negotiator. (Symposium on Cloud Computing, Santa Clara, California 2013).

[6] Zhang, M., Zhang, R., Liu, C.: Design of Smart Healthcare Data Management System Based on Hadoop. submitted to Advanced Materials Research (2014).

[7] G. Kousiouris, G. Vafiadis, T. Varvarigou: A Front-end, Hadoop-based Data Management Service for Efficient Federated Clouds. (IEEE International Conference on Cloud Computing Technology & Science, Athens, Greece 2011)

[8] Zhang, S., Han, J., Liu, Z., Wang, K.: Research on Spatial Query Based on MapReduce (in Chinese). Chinese High Technology Letters. 20(7), 719–726 (2010)