International Journal of Grid Distribution Computing
Vol. 8, No.5, (2015), pp.47-54
http://dx.doi.org/10.14257/ijgdc.2015.8.5.04

Design of Electric Energy Acquisition System on Hadoop

Yi Wu[1] and Jianjun Zhou[2]

[1]*School of Information Science and Technology, Heilongjiang University Harbin, Heilongjiang, 150080, China*
[2]*School of Information Science and Technology, Heilongjiang University Harbin, Heilongjiang, 150080, China*
[1]*wy51cn@aliyun.com*, [2]*zhou_1969@tom.com*

Abstract

The big data set in Energy Acquisition System needs to acquire massive electric energy data and dynamic information online, and finishes processing in scheduled time. This requires a higher demand on massive data storage and data processing. In order to achieve these massive electric energy data efficiently, this article based on the data gathering system and storage structure of Hadoop technique, and tested electric energy mensuration log data set of a city as an example, the result shows that the bigger of the sets group, the better effect would be achieved, which effectively avoid the latency problem of big data set information processing respond.

Keywords*: Electric Energy acquisition, Distributed structure, Hadoop calculation, parallel processing*

1. Introduction

Electric energy acquisition is the important element of power system marketization, upon completion of electric energy acquisition system, it will improve the level of electric energy information on real-time collection and supervisory control, accelerate the modernization construction of electric energy marketing, boost the innovation of marketing management pattern, enhance the ability of marketing management and service and achieve the goals on marketing modernization construction of the company, which are intensification, delicacy management, operating cost reduction and company's profitability improvement. The electric energy acquisition fits the need of sustainable development. It also turns the development direction of electric energy acquisition on centralization, automation and remote. There are four physical constructions of electric energy acquisition, master station layer, turn layer data set, data acquisition layer and customer electric layer. Meanwhile, for the purpose of building decision analysis system, it is necessary to take the best of active database from each relevant department, adopt the collection data of Harbin City as measuration and analyze data processing performance of the system.

As the development of internet and cloud computing technology, as the arrival of big data era, the electric energy collection data will increase sharply under the environment of intelligence network; it requires better reliability and instantaneity, which far more exceeded the processing scope of traditional grid electric energy monitoring [2]. In China, most of the construction of traditional information platforms for power industry adopts the expensive giant servers, uses disk array storage; database uses relational database system; operational application adopts tightly coupled suit which lead to bad system expendability, high operating cost and hardly adapt the necessary of reliability and instantaneity on intelligence network electric energy acquisition. Besides, traditional database call pattern changed to customer/server pattern, which first connect database with main program, and then obtain data through SQL operation, disconnect database in

ISSN: 2005-4262 IJGDC

the end. When it comes to call request, the traditional data call pattern must reconnect with database. Because of the instantaneity requirement of electric energy acquisition, even in a short period, the call requests would be hundreds of times. When the system is abnormal in processing, it will produce a large number of sporadic real time data, which may arouse to a mass of data call requests, lead to frequent database connection and disconnection. It results highly large system overhead. Therefore, when it comes to call requests processing with big data set information storage, traditional database call pattern could not satisfied the actual application requirements [3].

Recently, the environment of high performance distributed network parallel computation, represented by Hadoop architecture, is deployed step by step [4]. It brings resources like high performance communication, computation and storage, which provides unprecedented opportunity for large volume data of electric energy acquisition. Under the Hadoop architecture, see Figure 1. For electric energy data acquisition pattern design, by dividing the big data set information of Hadoop architecture into several small data chunks, it then equitable distributes to inside the cluster and parallel processed from the nodes. It could effectively prevent the multiple call requests from traditional database pattern, which greatly reduced the system overhead [5].

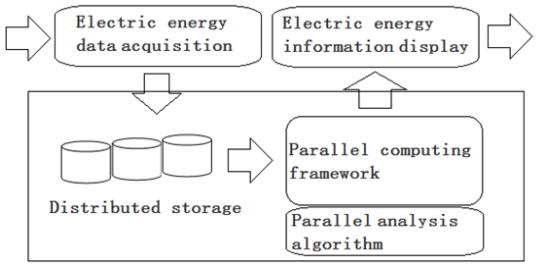

Figure 1. Electric Energy Data Acquisition Model

2. The Design of Electric Energy Acquisition System

Hadoop is the processing architecture which is facing large scale data processing; it could run on multiple system platforms. Besides, it has favorable reliability and extendibility with the advantages like large volume of data processing, high instantaneity and low cost. Hadoop aroused attention from lots of internet companies since it was published. The distributed architecture of Hadoop could easily process the massive unconstructed data in cyberspace. Hadoop Distribute File System (HDFS) is the distributed file system of Hadoop architecture. HDFS has the character like high fault tolerance. It uses master/slave structure which is convenient for designing and deploying on cheap hardware. HDFS could provide high throughput data access and process, which satisfied the demand of large volume data processing. Under the architecture of Hadoop, Map Reduce programming pattern is a distributed parallel computing model on processing large volume data. It applies to the parallel computation of large scale data. After introducing Hadoop parallel processing architecture, the bottleneck problem like the increasing of electric energy acquisition system data volume could be resolved.

2.1. Hadoop Rack Technology Foundation

Haddop is constituted by multiple elements. Hadoop Distributed File System (HDFS) is at the bottom. It saves all the files on storage node of Hadoop cluster. HDFS is a Master/Slave structure system. It constituted by NameNode and DataNode. Because of the character of distributed computation, at present, one Hadoop cluster includes one NameNode and thousands of DataNodes (See Figure 2.). Hadoop provides users the namespace of files and allows the storage of data to files.

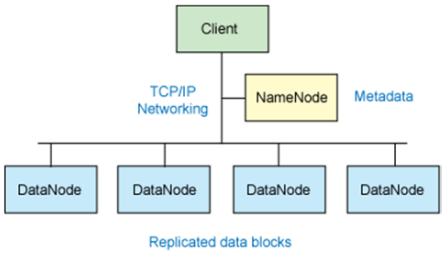

Figure 2. Hadoop Cluster Node Architecture

DataNode run to every machine in Hadoop cluster separately. DataNode is responsible on responding the data reading and writing request from HDFS clients. Meanwhile, it responses the instructions like create, copy and delete data files from NameNode. DataNode sends heartbeat message to NameNode at fixed period; each message includes one block report which could guarantee the supervisory control from NameNode to DataNode. NameNode could verify the accuracy of block mapping and other metadata in file system from the received block report. If DataNode could not send heartbeat message, NameNode would re-replicate the losing data block at this node [6].

Hadoop Distributed File System (HDFS) possess high fault-tolerant, which suits to deploy on the cheap hardware like common personal computers. It would achieve the accessing file system data in the form of stream which makes the data access in each node as the foundation of distributed system. Besides, it backed-up the data with several copies which avoid the possibility of losing data. As a result, HDFS has a relative mature fault handling mechanism. Figure 3 is a structure diagram of HDFS [7].

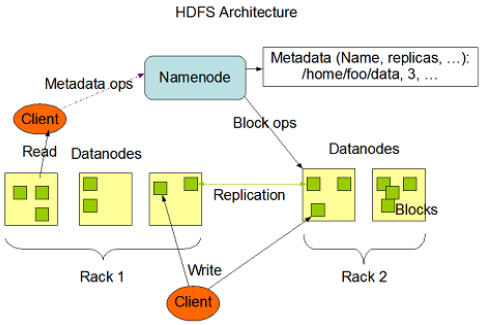

Figure 3. HDFS Architecture

The upper layer of HDFS is MapReduce programming model, which could achieve the processing of big data and adopt extensively. MapReduce is constituted by JobTrackers and Task Trackers. The user assigns one map function operates one key/value pair, and then assigns one reduce function which combined all the middle value with same middle key. This is the method which could achieve parallelization in large scale of common machine automatically [8].

MapReduce could process in flexible and adaptable cluster which is constituted by common personal computers. One typical MapReduce computation could handle data based on TB in thousands of computers. Map/Reduce is a easy-to-use software in Hadoop. The application based on it could run in large cluster with thousands of business machines, and parallelly process the T level data set with a reliable fault-tolerant way. One Map/Reduce job could divide the input data set into several independent data blocks, and map task processes them with completely parallel way. Structure could put the map output into order, and then reduce task do the result input. In general, the input and output of the task would be saved in the file system. The whole rack is responsible on scheduling

and monitoring the tasks and re-performing the failed task. Figure 4 illustrates the procedures of big data set process by using MapReduce [9].

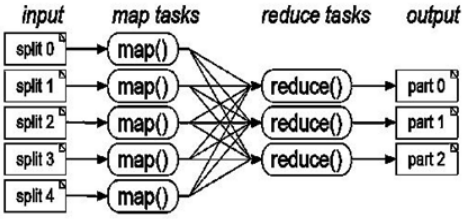

Figure 4. MapReduce Data Processing Model

Hadoop Database (HBase) is a high reliable, high performance, column oriented, scalable distributed database storage system. The HBase technology could build large scale structuring storage cluster on cheap Personal Computer Server. HBase make use of HDFS as its file storage system, by processing MapReduce in massive data of HBase, it could provide random and real-time read-and-access of large data. HBase is a distributed, column oriented model store loose data [10].

Based on above technique, by using the existing machine to build distributed, higher performance, high reliable and real-time electric energy information data processing and storage system, possess the condition of building low cost, high efficient electric energy information acquisition data analyzing system. This article aims at the increasing large scale electric energy data in powered grid, designed and realized the data processing and storage system on Hadoop platform.

2.2. Hadoop Architecture Technology Foundation

Based on electric energy information acquisition system architecture of Hadoop, there are four layers, storage layer, calculation layer, control layer and application layer. See Figure 5 for system architecture.

Functional design description:

(1) Storage layer

The storage layer saved all the data acquired from electric energy information, including historical data in database, condition-based inspecting data and metadata. Due to historical database, first, by using the import and export tool of Sqoop data, original data is extracted and uploaded to the storage layer, and then, HDFS read the data which is satisfied by the necessary of upper computing task. After completing implemented inquiry, compute and analyze tasks, the result could be derived to external condition monitoring historical database system, the imported original data could generate aggregate data which could be saved in the storage layer or directly delete by requirement. Due to the constantly producing of electric energy acquisition data, it would be stored in data files of the storage layer and HDFS would read the computation task processing to upper layer. HDFS would acquire the latest generated electric energy data file in a fixed period, and update the electric energy data set on HDFS. Metadata is generated in the process of building data warehouse. They are all saved in database.

(2) Calculation layer

The calculation layer uses the MapReduce programming model of Hadoop as the method of calculating electric energy information. The calculation tasks of Electric energy acquisition information including historical data search, multidimensional analysis, report generation, incremental maintenance, metadata access *etc*. The calculation

layer read and writes different type of electric energy acquisition data in storage by implementing control layer. The calculation layer could break apart any calculation tasks into two kinds of tasks (Map and Reduce), which dynamically distribute to different nodes and implement in cluster.

(3) Control Layer

The control layer includes the database engine which is composed with two query languages, HiveQL and SQL. The database engine deal with different requests from application layer, producing calculation tasks to calculation layer. HiveQL is used for analyzing query sentence in data warehouse, the requests from application layer transfer to Hive sentence in the control layer, by way of HiveQL analyzing, generage MapReduce task and invoke the calculation layer to perform, the produced result returns to the client through Hive user interface. SQL is used to manage the metadata information in Hive architecture data warehouse. Though the Hive table definitions, fields and space mark information which is created by HiveQL, all of these information would be save in MySQL relational database. On implementing the data operation of data warehouse, to begin with, launch the SQL engine to assure the existence of metadata.

(4) Application Layer

The application layers mainly constitute the function units like assistant decision and condition monitoring for electric energy information acquisition. It achieves the functions like search, calculation, analysis and decision of status information. The achievements of these unit functions are relied on controlling massive data. The support of functions in control layer is needed. Therefore, it would be easier to switch in application layer. The application layer also provided a serious of user interface, which is convenient for user access file system, user requests submission and data warehouse management.

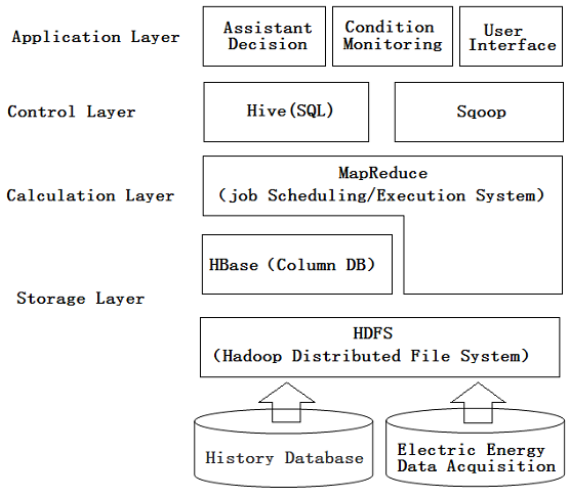

Figure 5. Information Acquisition System Architecture of Electric Energy

3. System Testing and Analysis

Electric energy data includes statics historical data and dynamic real-time acquisition data. As the time increasing, dynamic data needs more and more storage space and gradually exceed the storage ability of traditional data type of storage. At the same time, as the increasing of data amount, the processing ability of data is decreasing. In order to solve the current large scale data processing bottleneck, this system is based on achieving Hadoop, following MapReduce programing model and constituting by one master (NameNode) and multiple slaver (DataNode) nodes to parallel cluster storage and reading

pattern. NameNode is a master server, which is in charge of managing the file system with name space and the storage of all node data with clients. It also burdened on the reading requests from file system clients, as shown in Figure 6.

The system data is from Harbin power grid electric energy data acquisition system, including time scale, user ID, channel number, power supply bureau number, meter reading, electric energy of positive active power, data state, electric meter mailing address, rates *etc*. The total data size reaches 1.3 billion records.

First, launch Hadoop architecture calculation service platform in master scheduling machine. The platform contains Master/Slave structure composed of NameNode and DataNode, and the tasks deployed by JobTracker and Tasktracker in MapReduce Model. Among which, NameDode as the monitoring system data processing status of master mainframe, DataNode as processing electric energy data of Slave machine. Adopting MapReduce model to realize electric energy data parallel processing, the procedure code uses Eclipse (Java virtualized environment) to realize programming. Cluster platform structure has five machines; the internet is gigabit Ethernet, Hadoop version 1.1.2.

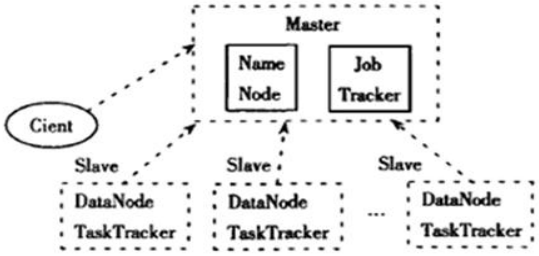

Figure 6. Hadoop Cluster Node Architecture

After that, launching the request data process of tasks invoking, according to decomposition rule, cluster NameNode processing information tasks segmentally through map tasks, deliver to DataNode to parallel calculation processing, and then hand over to Reduce tasks which process the final merger, aggregation and saved to distributed file system HDFS.

3.1. Build Experiment Platform

In order to test and verify the distributed storage invoking the applications of electric energy information acquisition system which is built by Hadoop calculation Architecture, combined the demand of large scale electric energy data set information processing, built one master mainframe and four task monitoring data slave, the constitution of HDFS distributed dispatching monitor cluster environment, as shown in Table 1.

Table 1. Distributed Dispatching Monitor Cluster Environment

Cluster	Computer configuration	Operating system	Number
Namenode (Master Machine)	Intel Xeon CPU E3-1235, 3.20GHz, 8M Cache, 5.0GB	Linux Ubuntu 13.04	1
Datanode (Slave Machine)	Intel CORE CPU I5-2400(vPro) 3.1G,6M Cache, 4.0GB	Linux Ubuntu 13.04	4

Based on the existing resources, installed domestic Linux System on Hadoop cluster, composed internal network (NameNode, DataNode1 to 4) and set up JDK, Hadoop and HBase on every machine.

3.2. Testing Analysis

The calculation of Hadoop and MapReduce replies on the HDFS distributed cluster Architecture, after launching the cluster environment, calculation engine processing electric energy big data set by using parallel mode. Under the traditional database mode, single node mode and cluster mode, choosing the record measuring data respectively are one million, three millions, eight millions, twelve millions and thirty millions for invoking access tests. The curve of the system's implementation time shows in Figure 7. From Figure 7, when invoking twelve millions measuring data, the implementation efficiencies are relatively equal with three modes; when increasing the measuring data to thirty million, under the cluster mode, the processing efficiency highlight the advantage, when the login history reached thirty millions, the reaction time is around 32.208s, which is 23% faster than traditional mode, the faster efficient processing, meanwhile, the whole reaction time curve shows a gentle shape. It verifies that the Hadoop cluster calculation technology suits parallel processing of large scale data size.

In addition, under the different scope of cluster mode, choosing the number of cluster 1, 2, 3, 4, 5, recording the data by invoking call test with the measuring data with eight millions and thirty millions, the speed up ratio curve of the system shows in Figure 8, which proves that the Hadoop cluster calculation technology suits parallel processing of large scale data size.

4. Conclusion

Due to the limitation problem of big data set processing efficiency under traditional client/server mode in power grid electric energy acquisition system, brings a method which is based on distributed cluster calculation method of Hadoop. According to the experimental test on different measuring data set of Harbin electric energy acquisition data, results indicate that comparing to traditional database mode and existing method, the processing efficiency of the new method is more effective and reliable.

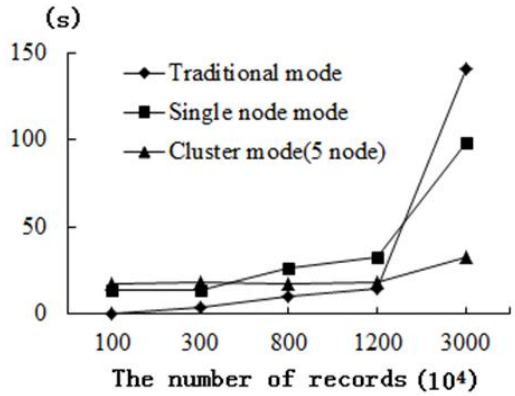

Figure 7. Time Contrast for Compression with Different Measurement Data Sets

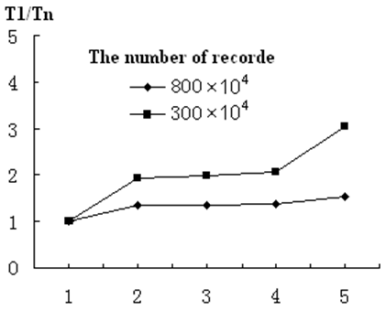

Figure 8. Performance Test for Speed-Up Ratio

Acknowledgment

This work is supported by the Nature Science Foundation of Heilongjiang Province with the grant number: F201325. The author would like to thank for their support.

References

[1] W. Yi, "Design of Energy Billing Database System Based on Data Warehouse", Journal of Natural Science of Heilongjiang University, vol. 2, no. 20, **(2003)**, pp. 66-68.
[2] H. Suleiman, K. A. Ahmed, N. Zafar, *et al*., "Inter-domain analysis of smart grid domain dependencies using domain-link matrices", IEEE Transactions on Smart Grid, vol. 3, no. 2, **(2012)**, pp. 692-709.
[3] Y. Sawano, "Generalized Morrey spaces for non-doubling measures nonlinear differ", Equ Appl, vol. 15, no. 4, **(2008)**, pp. 412-425.
[4] T. T. Zheng and S. Y. Zhang, "Boundedness of a class of operators over generalized morrey spaces with non-doubling measures", Journal of Zhejiang University of Science and Technology, vol. 23, **(2012)**, pp. 01-05.
[5] G. Hu, H. Lin and D. C. Yang, "Marcinkiewicz integrals with non-doubling measures", Integral Equation and Operator, vol. 58, **(2007)**, pp. 205-238.
[6] W. Feng, Q. Jie, J. Yan, *et al*, "Hadoop high availability through metadata replication", New York, Suite 701 New York NY USA, **(2009)**, pp. 114-133.
[7] K. Zheng and Y. Fu, "Research on Vector Spatial Data Storage Schema Based on Hadoop Platform", International Journal of Database Theory and Application, vol. 6, no. 5, **(2013)**, pp.85-94.
[8] G. Zhao, "A Query Processing Framework based on Hadoop 261", International Journal of Database Theory and Application, vol. 7, no. 4, **(2014)**, pp. 261-272 .
[9] K. Ullah and M. N. A. Khan, "Security and Privacy Issues in Cloud Computing Environment, A Survey Paper", International Journal of Grid and Distributed Computing, vol. 7, no. 2, **(2014)**, pp. 89-98.
[10] K. H. K. Reddy, D. S. Roy and M. R. Patra, "A Comprehensive Performance Tuning Scheduling Framework for Computational Desktop Grid", International Journal of Grid and Distributed Computing vol. 7, no. 1, **(2014)**, pp. 149 -168.

Authors

Wu Yi, born in 1963, professor, his current research interests include large scale data processing, data mining, sensor network, etc.

