THE STUDY OF HADOOP-BASED ARCHITECTURE FOR POWER QUALITY MONITORING CLOUD MODEL

Minghao AI
China
ai_minghao@126.com

Linhai QI
China
qilinhai@ncepu.edu.cn

ABSTRACT

The proposition of Power Quality Intelligent Information System provide an effective platform for the monitoring, analysis and processing to all kinds of power quality problems in Smart Grid.However, with the acceleration of Smart Grid construction,continue increasing of the amount of power quality monitoring sites and gradually improving of the system,various types of monitoring and calculating data increased dramatically,showed a trend of big data. Power quality Big Data will become the greatest obstacle of the efficiency and instantaneity of power quality analysis, processing and data mining,and the low efficiency system will also make it more and more unbearable to the users. In view of the power quality Big Data, we need to seek a new method of calculation. Cloud computing is a computing model which can distributed computing tasks into resource pool make up of a large number of computers to parallel execution. For Big Data computing, cloud computing platform has its inherent advantages,which can integrate computing resources and greatly enhance the computational efficiency.This paper presents and initially implemented a Hadoop-based cloud computing architecture for power quality cloud model,whose aim is to integrate traditional computing and storage resources in power grid to solve power quality big data storage and computing problems. The model uses hierarchical structures and based on natural stratification grid to set up a hierarchical power quality cloud computing platform on the data center of power grid at the provincial, city and county levels. The overall uses SOA architecture and encapsulates each layer and system in the form of service,and we should develop a common service bus. The layers build a sub-cloud based on Hadoop architecture.Each layer provides services to the upper main cloud. Through HDFS and HBase distributed data storage technology and MapReduce parallel computing technology in Hadoop,the power quality big data storage and analysis methods of parallelization are implemented,which can improve calculation efficiency and user experience. Hadoop-based architecture for power quality cloud model can not only solve power quality monitoring big data storage and computing problems,but also have the reference value to solve other information systems' big data problems in Smart Grid. .

FOREWORD

With the acceleration of Smart Grid construction, the information system becomes more and more important. The proposition of **Power Quality Intelligent Information System(PQIIS) provide an effective platform for the monitoring, analysis and processing to all kinds of power quality problems in the power grid.** However, As the continue increasing of the amount of power quality monitoring sites and gradual improving of the system, **various types of monitoring and calculating data increased dramatically, showed a trend of Big Data. Cloud Computing has inherent advantage to deal Big Data** because of the effectiveness of resource allocation and powerful ability on data processing.

Cloud Computing is a computing model which can distribute computing tasks into resource pool consist of a number of computers and provide computing ability, storage space and information service for users by need. [1]

Hadoop is an open-source architecture of Cloud Computing, which provide solution for analysing big data. It's key technology includes HDFS, MapReduce and HBase. **HDFS** is an efficient and large-scale distributed file system. MapReduce is a paiallel calculate model. **HBase** is a distributed database that support unstructured data storage.

The paper proposes a Hadoop-based architecture for power quality monitoring cloud model, builds power quality cloud platform based on **IaaS** and **PaaS**, and realizes layered power quality cloud structure that provide service in form of **SaaS**. The model provide valid solution for the collection, analysis and concentrate of power quality data.

OVERVIEW

Overview of power quality monitoring

As the proposition of Smart Grid, the need of transformation consist of grid structure and load as well as power users turn to higher reliability, more excellent power supply and more reasonable power price[2][3]. The noun Power quality takes the attention of people again. Improving the power quality to ensure the steady working of power grid has become one of the significant aims of Smart Grid.

The reference[4] gives reference definition of power quality problem: each electrical problem that shows the voltage, current or frequency deviation and causes the users' equipment fault or wrong action is power quality problem.

PQIIS is a platform that integrates power quality data acquisition, storage, index evaluation, problem analysis, comprehensive treatment ,data display and other functions. However, in recent years, PQIIS gradually feels overwhelmed on data storage and processing with the improvement of the function and running around. Detailed performances as follows:

The data is volume. Along with the increasing number of monitoring sites, power quality data increased exponentially. The original database storage pressure surge. To deal large number of data, we have to use higher performance hardware than now.

The data is seriously heterogeneous. First, the data's formats are various. Power quality data includes the sine waveforms under the rated voltage and rated frequency as well as all the transient phenomenon, such as shock pulse, damping oscillation, transient interruption and trap, etc. Its types include waveform data, event data, image data, geographic information data, and so on. Second,the data distributes in different regions, so the whole systemneeds to deploy regional.

The low density of data's value. In the case of sag event analysis. Sag events not happen very regular, but we will continuously record the voltage waveform data, in case of recording and analysis of waveform before and after the sag event occurs. For the whole waveform tens of thousands of points, the data useful to us may be just a few or dozens of points. So the valuable data density is very low.

Strong data real-time requirements. For some functions, such as pre-alarm, SVG display, have great demand of real time monitoring data.This requires us to adopt a technology for real-time data rapidly acquisition, transmission and processing.

The degree of power quality data mining is not deep enough. In view of the huge amounts of power quality data, the traditional mining algorithms is not applicable, so we must seek a new mining algorithm to a deeper and wider analysis data.

To sum up, **the power quality monitoring has entered the era of Big Data.** So we need to seek the way to solve the problem of big data, to improve the existing monitoring system architecture, and to use a new technology to store and handle this kind of big data, which is the cloud computing technology.

Overview of cloud computing

Cloud computing is the product of development and integration of distributed computing, parallel computing, utility computing, network storage, virtualization, load balance and other traditional computer and network technology. Cloud computing includes three levels of service which are Infrastructure as a Service(IaaS), Platform as a Service(PaaS) and Software as a service(SaaS)[1][5]. Google, IBM, Microsoft and Amazon on abroad as well as Alibaba, Baidu, 360 and lenovo in the domestic develop their company's cloud computing industry, set up their own cloud platform. At present, there are serveral mature cloud platforms such as Google's cloud computing search engine, Microsoft Azure cloud platform, amazon's EC2 elastic cloud platform, the IBM LanYun, Alibaba's ali cloud, etc.

The application of cloud computing platform in electric power industry is now in the research stage. The reference[6-8] proposed how to build power system cloud computing platform in the smart grid, and how to to seek a way to integrate the existing computing and storage resources in view of the traditional power system computing platform's shortage in computing, storage, information integration and analysis etc. The realization of the power system cloud computing platform in terms of physical composition, main function, system architecture and key technology was introduced. The reference[9-12] expounded **the application of cloud computing** in the aspect of computing and storage **in power system**. The reference[9] proposed the **storage of massive wave record data** by using Hadoop method and proved the efficiency compared with the traditional way. The reference[10] not only proposed **the equipment state monitoring data** distributed storage method based on Hadoop, but also implemented the data storage system, included Hadoop cluster, the storage client and query client. The reference[11] aimed at the shortage of huge amounts of state data's query and analysis in existing electric power data warehouse, proposed **the power equipment state information data warehouse** and **multidimensional data query and analysis method based on Hive.** The reference[12] used the relational database combined with Hadoop cloud computing platform, thus solving the the relational database's storage and access efficiency problem in the transmission line condition monitoring system.

PROPOSITION AND IMPLEMENTATION OF THE POWER QUALITY MONITORING CLOUD MODEL

Power quality cloud model

Power quality cloud model will use the **hierarchical** structure, be based on the electric grid natural stratification and deploy cloud platform at the provincial, city and county levels grids' data center. The the overall is **SOA architecture**. The **sub-cloud will provide data integration, interface calls and other services** to its higher level main cloud. The overall logical architecture of the power quality cloud model is shown in figure 1.

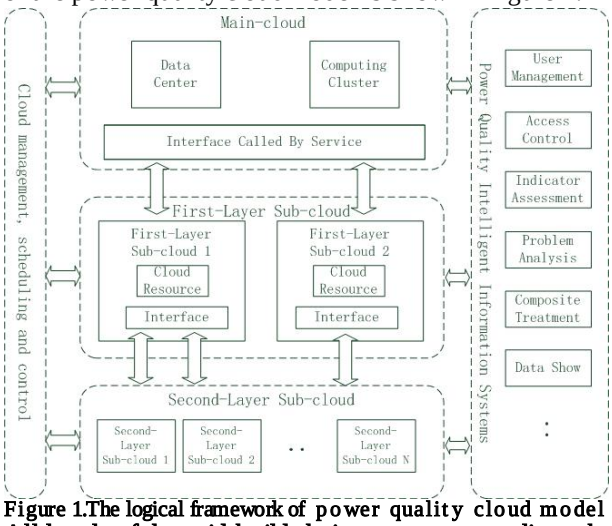

Figure 1.The logical framework of power quality cloud model

All levels of the grid build their own power quality sub-cloud. The existing various computing and storage resources are integrated to constitute the computing cluster by using **virtualization** and **segmentation** technology. The sub-clouds use the **distributed file system**(DFS) and **distributed database** to store the corresponding levels' cloud data on the cluster. The logical architecture of each level of the sub-cloud is

shown in figure 2.

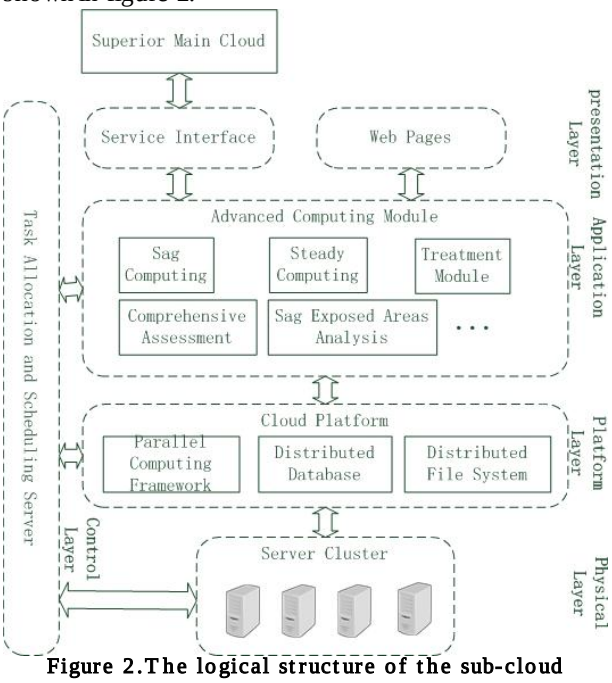

Figure 2.The logical structure of the sub-cloud

The implementation of power quality cloud

Overview of the implementation

In this paper, the power quality cloud model is **a layered cloud model,** which will set up layers of cloud platform respectively based on all levels of the grid. Calculating and data will be called **in the form of services among each layer**. That is to say, **the overall cloud model will adopt service-oriented architecture, and each layer's model will use Hadoop to build cloud computing platform.**

The overall SOA architecture

Service-oriented architecture (SOA) is a method of architecture which consists of a series of services interact with each other. SOA shields the differences between different platforms, programming languages, operating systems and hardware architectures, and realizes higher reusability and flexibility as well as better scalability and availability of services.

As the **regional distribution** and **platform heterogeneity** of all levels of the grid, the power quality cloud model built on SOA architecture has the advantage of transparency to layer. SOA architecture is divided into service providers, service bus and service consumers. One sub-cloud is service provider to its upper layer main-cloud, but it is service consumer to its lower sub-cloud. To various types of sub-clouds as services, we should develop a service bus for the main-cloud in order to make sure the transparency. Services can be encapsulated by **Xfire** method[13]. The schematic of using SOA to achieve the overall structure of power quality cloud model is shown in figure 3.

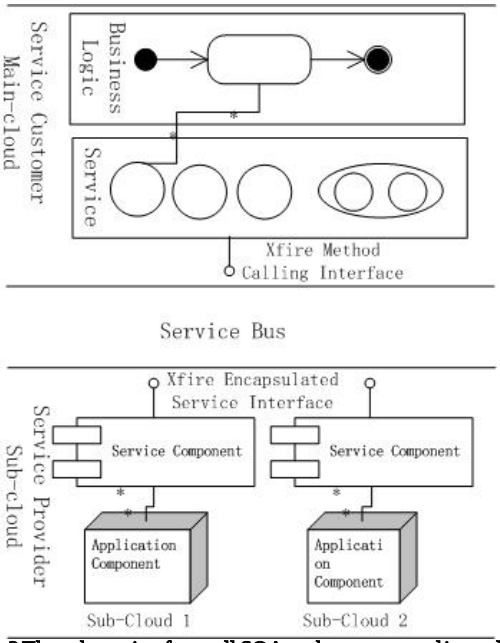

Figure 3.The schematic of overall SOA to the power quality cloud

Every parts in the overall SOA structure will all be implemented by using **cloud computing.** First, in underlying hardware, data layer and business layer, power quality basic application will be deployed in the form of private cloud(**IaaS and PaaS**), so that to improve the utilization of system resources. Second, in service layer, **using web service method to encapsulate each function distributed in the clouds.** The last, in presentation layer, in the form **SaaS**, enterprise applications can be integrated perfectly by using a uniform platform. These can enhance the user experience.

Each layer sub-cloud's Hadoop architecture

Hadoop is Apache's open source distributed computing platform. We can easily create and use distributed computing platform by using Hadoop as well as easily develop and run massive data processing application program on it. Hadoop has the features of high efficiency, high reliability, high scalability and fault tolerance. The main work to build layers of power quality sub-cloud architecture is to build Hadoop platform on clusters, to deploy distributed file systems and distributed database and to deploy the parallel implementation of the power quality analysis algorithm on the distributed platform.

(1) Building Hadoop platform on clusters.

There need a **namenode** pre-installed any version of Linux operating system and a **cluster** without requirement for operating system. But any node in the cluster needs the java development environment and the Secure Shell(SSH). The whole cluster requires sharing a network switch with at least 1Gb of bandwidth, so that the network would not become the bottleneck of computing efficiency. After determining the LAN cluster, the cluster topology structure will be defined according to the LAN. Using the tree structure to describe the topological structure, so as to master the location of each node and provide the basis for distribution of parallel tasks. Hadoop should be

configured to fully distributed mode. Then we should configure the files of core-site.xml, hdfs-site.xml, mapred-site.xml and mapred-queues.xml under the path of hadoop/conf according to the cluster topology and the IP address. Hadoop distributed file system(HDFS) has been integrated in Hadoop after configuration is complete. Before the first time starting the hadoop we need to format the HDFS file system, and then start hadoop by start-all.sh script and test the connectivity between nodes.

(2) Deploying the distributed database.

HBase is a column-based stored distributed non-relational database in Hadoop project. The query efficiency of HBase is so high that the query result can be displayed in time. **Hive** is a distributed data warehouse, which is mainly used in parallel distributed processing large amounts of data. Hive provides a mechanisms to store, query and analysis large-scale data stored in Hadoop, and defines a simple SQL-like query language called HQL. Since HBase has no SQL like query mode, and it is very inconvenient to manipulate and calculate the data, integrating HBase and Hive should be considered, which **use HBase as the bottom storage and Hive as the support of HQL queries** in the database level of HBase. We use **MySQL** database to manage Hive metadata.

The implementation of integration of HBase and Hive is **mainly using their own external APIs to communicate with each other which rely on hivehbase - handler. jar utility class.** Copy the jar package to the lib folder under Hive and HBase. At the same time, we need to configure the documents of hive-site.xml, hive-log4j.properties, hive-env.sh and hbase-site.xml.

(3) The parallelization of power quality analysis algorithm.

Algorithm parallelization is the core of cloud computing applications. Using the **MapReduce** programming model can greatly reduce the workload of parallel programming, because MapReduce provides a number of API and programmers just need to mainly implement its Map function and Reduce function. **Map function** receives an input form of <key,value>. Multiple Maps excute in parallel and each Map produces an intermediate output in the form of <key,value>. Hadoop is responsible for collecting all values with the same key and passing them to the Reduce function. **Reduce** receives a <key,list(value)> form input, processes this value list, and finally, produces a <key,value> form output. MapReduce operation mechanism is shown as in figure 4.[14]

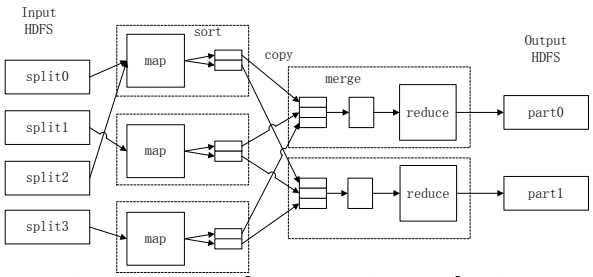

Figure 4. MapReduce operation mechanism

In the power quality analysis algorithms, there are many traditional algorithms can be parallelized. For example, in **harmonic calculation**, we can calculate each harmonic in parallel; in **sag analysis**, we can first send **data in different sag events, different voltage phases and different cycles** to various **Mappers to preprocessing** and then **collect the intermediate results** and **pass to Reducers to calculate the final result**; in **steady state index evaluation**, calculate each monitor sites, each days' data in parallel.

In addition to parallel the traditional power quality analysis algorithm, we can also **do some optimization in task level**. For example, the setting of **data preprocessing**, the setting of **secondary sorting** and the **compression processing of the intermediate results.**

(4) The release of the application system.

Choosing the **Java** language, application system is developed under the Eclipse IDE (integrated development environment) with the **Hadoop-Eclipse plugin**. Different database operation module and basic calculation module are encapsulated into jars for the calls from JSP front desk or Xfire encapsulated Web Service.

CONCLUSION

The paper first analyzed power quality monitoring data in Smart Grid **has the trend of Big Data** from various angles, and proposed the necessity of improving the method of calculation. Then it puts forward and initial implements a **power quality monitoring cloud model based on Hadoop framework.** The model used cloud computing platform that has the advantage of processing Big Data and built power quality monitoring cloud in the way of integration between SOA architecture and cloud computing. Based on this model, the power quality intelligent information system(PQIIS) was implemented. These **solved the problem of Big Data storage and computing in power quality monitoring** and have the **reference value to Big Data problems in other information systems in Smart Grid.**

The cloud model in this paper needs to be studied further in the following aspects.

(1) The existing power quality monitoring data is stored in the traditional relational database, but data storage in the cloud computing platform relies on the distributed database. How to **import the data from traditional relational database to the distributed database losslessly** will be will become a problem must be solved.

(2) The parallelization problem of traditional power quality data analysis algorithms. Most of the traditional algorithms are serial, in order to improve the execution efficiency of MapReduce, we need to improve the algorithms according to the MapReduce programming model. This needs to consider the problems of whether the algorithm itself can be parallelized and whether the efficiency can be improved after parallelization. Now, there is not a common method that can parallelize the traditional serial algorithm. So we have to **case by case**. The **parallelization problems of various traditional serial algorithms** will also become a new hotspot.

REFERENCES

[1] LIU Peng, 2010, *Cloud Computing(V2),* Publishing House of Electronic Industry, Beijing, China, 1-10.

[2] LIU Jun-yong,YANG Hong-geng,XIAO Xian-yong, 2004, "Issues and technology assessment on power quality Part 5:Power quality monitoring and data management", *Electric Power Automation Equipment,* vol. 2, 1-4.

[3] YAO Jian-gang, LI Ji-guang, LI Tang-bing, 2009, "Design and Appication of Novel Power Quality Monitoring and Evaluation System", *Proceedings of the Chinese Society of Universities,* vol. 2, 76-80.

[4] XIAO Xiang-ning, XU Yong-hai, 2001, "Power Quality Analysis and Its Development", *Power System Technology,* vol. 3, 66-69.

[5] ZHANG Kai, HE Ying, 2012, "Research on Power Simulation System Based on Cloud Computing", *Modern Electric Power,* vol. 6, 38-42.

[6] ZHAO Jun-hua, WEN Fu-shuan, XUE Yu-sheng, LIN Zhen-zhi, 2010, "Cloud Computing : Implementing an Essential Computing Platform for Future Power Systems", vol. 15, 1-8.

[7] YANG Xu-xin, LIU Jun-yong, JI Hong-liang, PAN Rui, HE Xing-qi, GUO Xiao-ming, 2010, "Electric Power Systems Cloud Computing Preliminary Study", *Sichuan Electric Power Technology,* vol. 3, 71-76.

[8] WANG Ai-min, YU Jin-gang, 2011, "The application research of cloud computing in smart grid", *Symposium on Chinese Smart Grid.Beijing*, China Electrotechnical Society, vol.1, 152-156.

[9] BAI Hong-wei, MA Zhi-wei, ZHU Yong-li, BAI Hong-hao, 2011, "The storage and access of the recorded data based on Hadoop", *Journal of the Hebei Academy of Sciences,* vol. 3, 43-47.

[10] LIU Shu-ren, SONG Ya-qi, ZHU Yong-li, WANG De-wen, 2013, "Research on Data Storage for Smart Grid Condition Monitoring Using Hadoop", *Computer Science,* vol. 1, 81-84.

[11] WANG De-wen, XIAO Kai, XIAO Lei, 2013, "Data warehouse of electric power equipment condition information based on Hive", *Power System Protection and Control,* vol. 9, 125-130.

[12] SONG Ya-qi, LIU Shu-ren, ZHU Yong-li, WANG De-wen, 2013, "Application of Cloud Computing Technology in the Transmission Line Condition Monitoring System", *Mathematics In Practice and Theory,* vol. 5, 109-115.

[13] LIANG Ai-hu, 2007, *SOA thought, technology and system integration application explain*, Publishing House Of Electronics Industry, Beijing, China, 28-29.

[14] LU Jia-heng, 2011, *Hadoop Action,* China Machine Press, Beijing, China, 56 .