Research on Parallel Processing Framework of Power Big Data

HU Bin[1, a], LUO Li-ming[1,b] , YANG Pei[1,c] , HUANG Tai-gui[2,d] and ZHANG Li-ping[3,e]

[1]Global Energy Interconnection Research Institute, Beijing, 102209, China

[2]State Grid Anhui Electric Power Company, Hefei, Anhui, 230061, China

[3]Nanjing University Post & Telecommunication, Nanjing 210003, China

[a]hubin@geiri.sgcc.com.cn, [b] luoliming@sgepri.sgcc.com.cn, [c]yangpei@geiri.sgcc.com.cn, [d]ds16090311@126.com, [e]295345439@qq.com

Keywords: Power big data, Parallel Processing, Power quality, Hive

Abstract. With the application of the power of big data landing and deepening, which put forward higher requirements on the performance analysis, and based on multi node parallel processing technology will become one of the core support technology and the rapid development of the application of power analysis of big data. This paper briefly analyzes the application requirements of large power parallel data processing, this paper presents a parallel processing framework for the power of big data and the realization method of management services and computing services in detail the design, research and development of parallel processing platform for the power of big data, and based on the national power quality monitoring data, comparison test Hive and 1.1.0. Compared with the experimental results, the results show that, when the node size is relatively small, the power big data parallel processing platform is less obvious than Hive 1.1.0. However, with the increasing size of the nodes, the advantages of the large data parallel processing platform of power are larger than that of Hive 1.1.0.

1 Introduction

With the continuous development and deepening of the smart grid, the coverage of the power terminal collection device is more and more widely, and the data acquisition frequency is higher and higher, which form an amount of power data resources. At present stage, the power big data has enter into maturity stage from the concept stage. The professional oriented large data analysis and application of electric power becomes a national Power Grid Corp foothold in practical application of power big data, which is urgently demanded in business analysis model from the data acquisition and analysis to the full amount of data analysis, data analysis from small batch data analysis to massive data analysis and from single business data analysis to related business data analysis and transformation, From offline data analysis to real-time / quasi real time data analysis, So as to fully meet the requirements of panoramic and full capacity high performance of data analysis and

application. Faced with the new requirements of the large data analysis and application, the multi node parallel processing technology is bound to become one of the core technologies to support the rapid development of the power data analysis application.

Parallel processing is one of the main ways to solve the large-scale and complex computing problems by using a variety of computing resources at the same time. Among them, high performance is mainly reflected in the use of a variety of computing resources to solve large-scale complex computing problems; then high concurrency is mainly reflected in the management of task decomposition and integration of computing results through concurrent work mode, At the same time, which simplifies the difficulty of parallel programming, and has the advantages of high throughput and low latency; however high availability is mainly reflected the dynamic scheduling and mechanism based on the resources, in order to improve the parallel computing task allocation mechanism and achieve the dynamic join of resource nodes and repair and exit of the actual node. Currently, Song *et al.*[1] proposed a method of parallel processing oriented data transmission equipment, and using the MapReduce parallel programming model, we design and implement the parallel query algorithm for multi data sources and the parallel feature extraction algorithm for multi-channel data fusion; Wang *et al.* [2] proposed a large data analysis and parallel load forecasting method for power users, and realized the parallel load forecasting system based on Hadoop; Qi *et al.* [3] design and implement a hierarchical parallel computing method of voltage sag based on the Hadoop cloud computing platform; Aiming at the massive data set of intelligent distribution network Qu *et al.* [4] and others use the Map/Reduce cloud computing engine to realize the distributed lossless cluster compression of the variable cross-section measurement data. The above research can be found in solving the problem of high performance computing more still uses the MapReduce framework, and the main research in parallel algorithm optimization direction, not on the parallel processing framework for further research and improvement. However, the most serious limitations of the MapReduce framework are scalability, resource utilization, and the support of different workloads in solving large-scale parallel computing problems.

Therefore, considering the new requirements of large data analysis and the shortcomings of the existing parallel processing framework, This article presents a power oriented big data parallel processing framework, and discusses a parallel processing method and optimization strategy, then researches and development of parallel processing platform for the power of big data, which effectively improve the efficiency of the power of big data parallel processing, and basically meet the power of big data panorama. It meets the requirements of large data analysis and application of panoramic, full and high performance.

2 Platform Framework

Basing on enterprise message bus (Enterprise Service Bus, ESB), this paper constructs the framework of large data parallel processing platform, which includes management services and computing services. It provides a common interface for parallel processing, as shown in Fig.1.

(1) Management services. Management service is the core of large data parallel processing platform, which usually contains one or more management nodes. The management node can communicate with the computing node to form an extensible computer cluster. Management services include task partition management, computing resource management, parallel configuration management and task scheduling management.

(2) Computing services. The computing service is the carrier of the large data parallel processing platform, which can be composed of several computing nodes and a single computing node is the basic unit to perform a specific task. Computing services include three modules: scheduling management, task execution management and task control management.

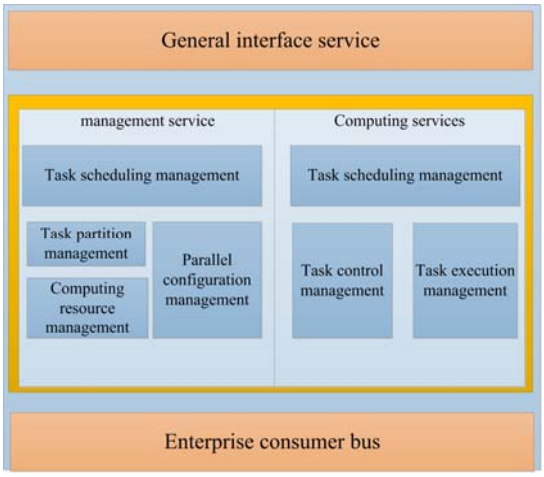

Fig.1. Power big data parallel processing platform framework

3 Platform logic

Based on the large data parallel processing platform, the parallel processing can be described as the following steps: task partitioning, task scheduling, task processing, and summary of results, as shown in Fig.2.

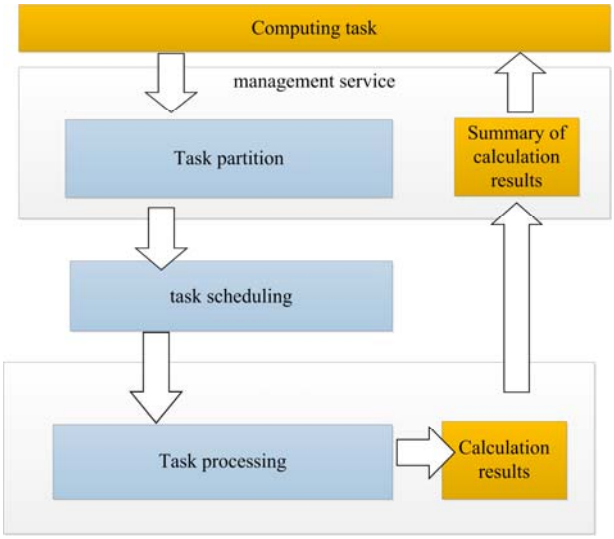

Fig .2. Power big data parallel processing platform logic

(1) Task partition. The computational task is a complex computing process initiated by the user, which is characterized by the large amount of data, high computing density and strong logic. Task partition is based on a certain business rules. The computing tasks are divided into a set of logical independent subtasks, which can further evaluate the complexity, priority and other parameters. After the task partition is completed, a list of tasks will be formed in the management service module for task scheduling.

(2) Task scheduling. Task scheduling is mainly responsible for scheduling the tasks, which commonly used scheduling algorithm[5] including First-Come First-Served(CFS), Shortest-Job-First, (SJF) Priority First-Fit, Heuristic algorithms(such as simulated annealing algorithm and genetic algorithm). The power of big data parallel processing platform with priority scheduling algorithm and priority parameters mainly based on computational tasks and computing nodes through the parallel

adaptive allocation method and task scheduling optimization. Among them, the priority of computing task is determined by task partition; the operation parameters of the computing nodes are collected by the platform resource monitor in real time, which includes the CPU, memory, I/O, storage and so on. The computing task is assigned to the computing node which can satisfy the requirement of the computing task.

(3) Task management. Task processing is mainly responsible for the specific process of the task, including task control and task execution. Among them, the task control is mainly aimed at the process control; the main task is to manage the process queue.

(4) Summary of results. The results define how to summarize the results of each computing task and other operations, which are mainly used to collect the results of each computing node.

4 Experimental analysis

The large data parallel processing platform and Hadoop platform, which consists of one main management nodes, one spare management nodes and seven computing nodes, are constructed respectively. The deployment architecture is shown in Fig.3.

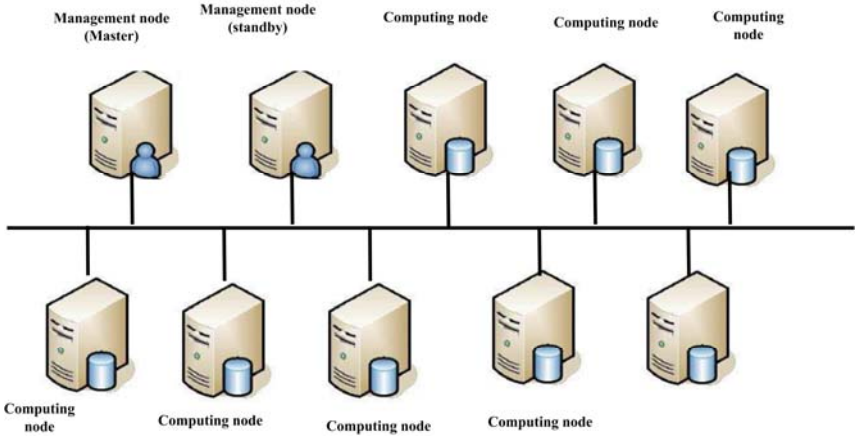

Fig.3. Deployment framework

In the deployment architecture, the management node and the computing node are unified by using the 2- way server. The specific configuration is listed in Table 1.

Table 1. Server configuration

Equipment	configuration parameter
Processor	Intel Xeon E5-4603, 2 channels and 16 cores
Memory	96 GB
Hard disk	2TB * 5 (SASA) , 7200RPM, 1TB*1(SAS)
Network card	Four port Gigabit Ethernet

The experimental data uses the national power quality monitoring data, involving a total of 7 data tables, about 130.00GB and the maximum number of single table records is 170 million Referring to Table 2.

Table 2 Experiment date

Data table name	Data sheet (GB)	Data record (bar)
Unit code	0.02	50,553
Medium voltage power outage across the lunar surface	0.01	11,062
Medium voltage blackout	4.53	8,068,309
The power line voltage across the lunar surface	0.01	27,130
Medium voltage power line	13.35	22,904,602
Medium voltage power users across the lunar surface	0.13	139,837
Medium voltage power user	109.97	170,000,000

Based on the historical data of the national power quality monitoring, this paper designs an experimental example of "medium voltage outage correlation analysis" across 7 data sheets. Its business logic is shown as follows.

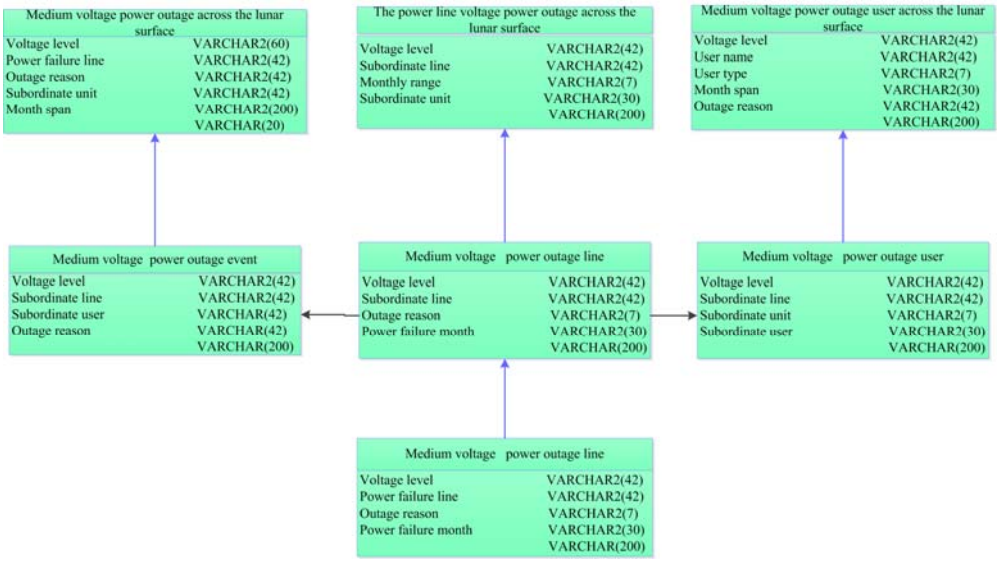

Fig .4. Experiment case

The experimental method is based on the power of big data parallel processing platform and Hive 1.1.0 in the 1 nodes, 3 computing nodes, 5 nodes, 7 nodes under the environment test. The results are shown as in Fig.5.

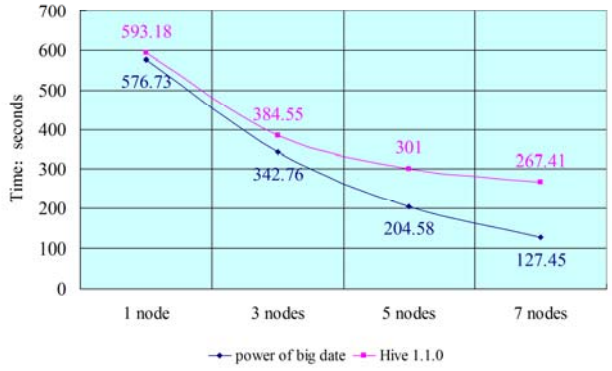

Fig .5. Experiment results comparing analysis

From Fig.5, we can see that in the case of relatively high degree of complexity of data processing, the node size is relatively small and the power big data parallel processing platform is less obvious than Hive 1.1.0. However, with the increasing size of the nodes, the advantages of large data parallel processing platform of power is bigger than that of Hive 1.1.0.

5 Conclusions

Multi node parallel processing technology will become one of the core support technology and the rapid development of electric power application based on large data analysis. In order to meet the application requirements of large power parallel data processing, this paper studies the development of parallel processing platform for the power of big data, and based on the same benchmark and Hive 1.1.0 are compared to test. The experimental results show that the parallel processing platform for large power data can effectively improve the efficiency of large data parallel processing, and meet the requirements of large data analysis and application of panoramic which is full and high performance. On the basis of the current research, the next step is to further optimize the methods and Strategies of large scale node scheduling, and improves the usability and practicability of the platform.

Acknowledgements

This work was financially supported by the State Grid Corporation of Science and Technology Projects (Research on Real Time Processing and Intelligent Analysis of Power Big Data).

References

[1] SONG Yaqi, ZHOU Guoliang, ZHU Yongli, LI Li, WANG Liuwang, WANG Dewen. Storage Optimization and Parallel Processing of Condition Monitoring Big Data of Transmission and Transforming Equipment Based on Cloud Platform. Proceedings of the CSEE, Vol.35 (2015), p.255-267.

[2] WANG Dewen, SUN ZhiweiBig. Data Analysis and Parallel Load Forecasting of Electric Power User Side. Proceedings of the CSEE, Vol. 35(2015), p. 527-537.

[3] Qi Linhai, Ai Minghao. A voltage sag parallel calculation method based on cloud computing. Proceedings of the CSEE, Vol. 34(2014), p.5493-5499.

[4] Qu Zhijian, Guo Liang, Liu Mingguang. New variable section flexible compression algorithm for measurement information in intelligent distribution network. Proceedings of the CSEE, Vol. 33(2013), p.191-199.
[5] LI Chuan-rong, CAI Xing-wen, HU Jian, HAN Ye-mao, LI Zi-yang. Design and Implementation of Universal Parallel Framework for Remote Sensing Process. Science Technology and Engineering, Vol. 35(2012), p.9540-9544.
[6] XIE Chen-ning. Distributed Graph-Parallel Framework Scheduling Analysis and Optimization. Shanghai Jiao Tong University, 2015.
[7] Ralf Starmark. Massively Parallel Processing of Recursive Multi-period Portfolio Models. European Journal of Operational Research, 2016.
[8] CUI Yang, LU Zhiping, CHEN Zhengseng, WANG Yupu, LU Hao. Research of Parallel Data Processing for GNSS Network Adjustment under Multi-core Environment. Acta Geodaetica et Cartographica Sinica, Vol. 42(2013), p.661-667.
[9] GU Guotai, XIAO Han. Application research of concurrent computing and concurrent Processing technique. Journal of Henan Polytechnic University (Natural Science), Vol. 5(2009), p. 621-625.