Large-Capacity Data Processing of Main Distribution Network Based on Information Processing Cluster Framework

Hongbo Wei[1](✉), Guinan Ye[1], Jiancheng Wei[1], and Hu Xie[2]

[1] Power Dispatch and Control Center of Guangxi Power Grid Co., Ltd., Nanning, Guangxi, China
xiaochongtougao5@163.com

[2] Digital Grid Research Institute, China Southern Power Grid, Guangzhou, Guangdong, China

Abstract. Due to the continuous development and in-depth promotion of smart grid construction in China, the amount of information accumulated has increased exponentially. The technical method of extracting "treasure" from these important historical materials has gradually become an urgent need for building a powerful intelligent power grid, and the rise of big data storage and processing technology has also opened up a new road for data mining. This paper studies the high-capacity data processing of main distribution network based on information processing cluster framework. After understanding the relevant theories, a high-capacity data processing system of main distribution network based on information processing cluster framework is designed and tested. According to the accuracy test results of the system, the accuracy of the system is about 70%, which basically meets the needs of the system, and then the system management efficiency is good. The effectiveness test results show that the parallel test time of the system is greatly reduced compared with the serial time, so the system has good parallel processing efficiency.

Keywords: Information processing · Large-capacity data · Data processing · Smart grid

1 Introductions

At present, the data of power companies is structured and semi-structured, and the growth rate changes from TB level to PB level. It has already bid farewell to the era of relatively uniform data types and relatively slow growth [1, 2]. Existing platform functions: data storage capabilities, data display capabilities, and data processing capabilities may not meet the data analysis needs [3, 4]. In addition, most data processing platforms now also use expensive mainframe computers, which have low scalability and high cost [5, 6], and disk arrays are mainly used to store a large amount of data, and one or several servers need to be upgraded. To reduce the storage time, these are not well enough for the processing and analysis of big data. Therefore, the data analysis platform must have good scalability and high fault tolerance. In order to provide better value-added services

I. Ahmad et al. (Eds.): STSIoT 2021, LNDECT 122, pp. 525–534, 2023.
https://doi.org/10.1007/978-981-19-3632-6_63

to our country's power companies, we can make full use of emerging technologies such as big data processing technology and massive data mining, comprehensively analyze most of the data resources, and find valuable information among them [7, 8].

Regarding the research of large-capacity data processing, some researchers pointed out that in recent years, data storage control systems have become more and more important, and storage testing technology is also constantly developing. High requirements are put forward on storage test technology and require test equipment reliability, high anti-interference, high data measurement accuracy, etc. [9]. Some researchers also pointed out that processing massive amounts of data is a major challenge facing all industries in this era, not only in the field of defense, but also in various fields of massive data. Data analysis can help companies make wise decisions based on this information, and decisions play a leading role in the rise and fall of enterprises. Therefore, research on massive data can bring great value to enterprises, which also promotes the continuous development and innovation of massive data processing technology, and drives the arrival of the era of big data [10]. Regarding the research on the information processing cluster framework, some scholars pointed out that there are still many problems in the information processing cluster framework. For example, JobTracker will increase the overall memory resources [11]. Some designers have used threads as a task management framework to solve the above problems, and redesigned the Map-Reduce framework with a new version of the information processing complex architecture, decomposing the old JobTracker into two independent components. One is dedicated to resource scheduling, and the other is dedicated to task control. In this way, tasks can be assigned to various resource filling states to reduce resource consumption on a single host and further optimize the resources on each node. This makes Hadoop a very large, powerful and reliable data processing framework [12]. In summary, there are still many research results on big data processing technology, but there are relatively few researches on active network data processing.

This paper studies the large-capacity data processing of the main distribution network based on the information processing cluster framework, summarizes the characteristics of the active network data on the basis of relevant documents, and then analyzes the application of the information processing cluster framework in data processing. Based on these, design a large-capacity data processing system for the main distribution network based on the information processing cluster framework, test the designed system, and draw relevant conclusions through the test results.

2 Research on Large-Capacity Data of Main Distribution Network

2.1 Main Distribution Network Data Characteristics

With the digital transformation of power generation intelligence, power transmission monitoring networks, smart substations, smart meters, etc., the scale and types of power data have increased dramatically, and huge power grids have generated a lot of big data. From power generation to electricity consumption, many data collection sources pay great attention to collecting specific data information, especially image formats. The final smart grid data is huge and diverse, the value density is very low, and the speed is very fast, so the characteristics of the main distribution network data are summarized as follows:

(1) Large amounts of data. With the construction of smart grids, grid data has increased from GB and TB to PB, EB, and ZB.
(2) There are many types of data. There are many types of smart grid big data, including structured, unstructured and semi-structured. With the growth of video applications, the proportion of unstructured multimedia data in the power network is gradually increasing. In addition, data applications require correlation analysis of various data types such as meteorological data and non-industrial energy data, adding data types and complicating data processing.
(3) High speed. The high-power data processing speed needs to be very high, which can reach the processing speed of microseconds, and can quickly analyze the data in a short time and support the database, so as to make accurate decisions on the reliable operation of the power grid. Compared with offline data, the requirements for online data processing are higher, and the analysis and extraction of Web data streams are very different from traditional data mining techniques.
(4) The value density is low. Take video viewing as an example, useful data is only 1–2 s. In addition, most of the information collected by the equipment is normal, with very few abnormalities. This happens to be an important basis for the operation and maintenance of the power grid.

2.2 Application of Information Processing Cluster Framework in Large-Capacity Data Processing of Main Distribution Network

(1) The distributed file system (HDFS) in the information processing cluster framework is designed to run on general-purpose hardware. Based on the cloud platform of the information processing cluster framework, the HDFS distributed file index is created, distributed massive data processing, and real-time search is provided. HDFS ensures that very large files can be stored on the machine, and each file can be stored as a series of data blocks. The files are written at the same time, and there can only be one writer at a time. This ensures reliable data storage. When the client receives the file, it checks whether it is suitable. Based on HDFS, large data blocks are decomposed into a large number of small data blocks with complete fault tolerance.
(2) The information processing cluster framework first sorts the distribution network data, then divides the map job according to the information processing module, processes multiple calculation tasks in parallel, and finally reduces the calculation results of each map. Effectively solve the problem of excessive data calculation delay caused by the excessive amount of existing data in the SQL database.

2.3 Data Processing Algorithm

(1) The MapReduce data processing model

The MapReduce data processing model is currently the most common data processing model in big data processing. Therefore, many big data processing systems based on the MapReduce data processing model have been implemented, such as Hadoop, Pig, and AsterData. In order to calculate and process large and small graph data, this section mainly introduces the Network TopValue PageRank algorithm.

The user randomly selects a specific network node (web page) as the default node. After the user initiates an access operation, the user randomly selects a directed edge to enter the next network node with a specific probability d according to the existing network graph G architecture (the probability d ranges from 0.1 to 0.2, usually 0.15). According to the above rules, in the directed graph G(V, E), the PageRank value of node u is obtained by formula (1). Where V represents the network node, and E represents the network connection (directed edge).

$$R(u) = (1 - d) + d \times \sum\nolimits_{v \in B} \frac{R(v)}{N} \tag{1}$$

where B is the set of all neighbors that have internal relations with node u, and N is the externality of different nodes v.

The PageRank value of each network node in Figure G is determined by the score, except for all nodes pointing to it and the corresponding PageRank value. If you want the final starting value of different nodes in the network, you need to repeat Eq. (1) until the starting value of each node does not change. The iterative expression is shown in (2).

$$R_i(u) = (1 - d) + d \times \sum\nolimits_{v \in B} \frac{R_{i-1}(v)}{N} \tag{2}$$

(2) Hadoop computing cost

The biggest advantage of the Hadoop big data processing system lies in the map mapping function of MapReduce, that is, the basic data processing model, that is, it can map input data to the local computer as much as possible. However, due to the limited hardware resources of each computing node, the local calculation of input data during the map processing stage also depends on the distribution of the data. If the number of data fragments on a particular node exceeds the maximum number of data segments that a node can handle, this means that some data nodes in the node must be used by other remote nodes, but cannot perform local map functions. Therefore, the number of map functions started by the data node in the processing phase of the map mapping function depends on the amount of data that the data node needs to edit. The cost of reading the Map function is shown in formula (3).

$$\mathrm{cost}M_{in} = \lambda * \overset{s}{D}_r + (1 - \lambda) * \underset{m}{\overset{s}{N}}_s + Cost_{map\text{-}calls}(num) \tag{3}$$

3 The Large-Capacity Data Processing System of the Main Distribution Network Based on the Information Processing Cluster Framework

3.1 Overall Framework of Large-Capacity Data Processing System

This paper proposes a large-capacity data processing system for the main distribution network based on the data characteristics of the main distribution network. The functional framework of the system is shown in Fig. 1, including data collection, data storage, data processing, and data management.

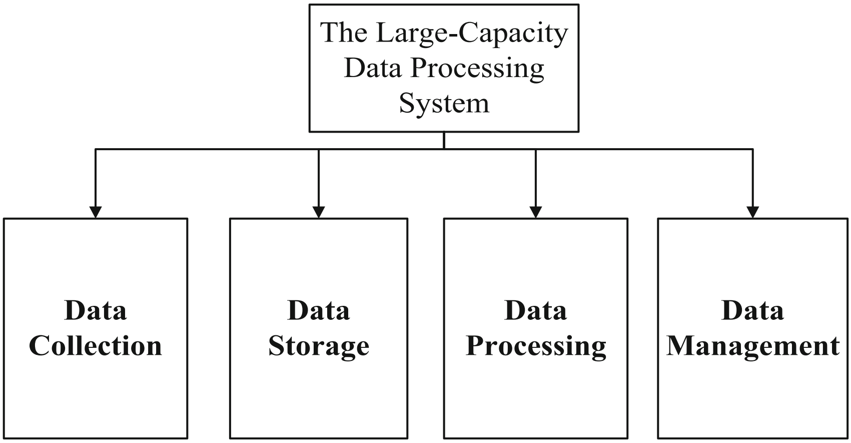

Fig. 1. The functional framework of the system

3.2 Data Collection

A large number of distribution network data are obtained from various monitoring terminals. According to the grid index system, it is divided into safety, reliability and power quality. The specific parameters mainly include power supply reliability, average number of power outages, distribution transformer operating life, switch operating life, average trunk length, etc. The monitoring terminal can select the appropriate collection frequency according to the characteristics of the index, generally once every 3 to 5 min. At this time, the monitoring terminal automatically collects data and chooses to send the data to the top layer for processing by the terminal. With the development of the power system and the increase in monitoring frequency requirements, the collection frequency will inevitably increase. Assuming that the acquisition frequency is set to the acquisition time of 2 min, each channel acquires 3 values, the amount of data collected per hour is 90, and the amount of data collected per day is 2160. Because each data has its own characteristics and different acquisition characteristics, some for data, such as harmonics, a sequence of at least 40 waves is collected.

3.3 Data Storage

At present, the power data warehouse can only meet the static statistical requirements, and the function is T+1. This means that big data in the power grid is statically stored, and its statistical data can only measure data in previous time units. The relationship of heterogeneous big data is complex, and information operators cannot build efficient big data warehouses above the XB level, which brings major problems to the data processing of information operators. Currently, it is mainly supported by various distributed technologies. According to the location of the data, the data will be distributed and stored in the local database. This mashup architecture is usually a parallel new-generation MPP+Hadoop database and some computing technologies. At present, this model can barely support, as the amount of data grows, this technology will soon become difficult to use. So this article applies new database technology.

The basic technology of the new database is very different from the traditional database. Efficiently and technically process PB data is to solve data storage problems for industry users. The new database will gradually integrate with Hadoop to provide rich SQL support for semi-structured and unstructured data processing. In this way, the processing needs of complex data are also met.

3.4 Data Processing

In order to standardize the complete production data chain of the system, this paper chooses to simulate an innovative model, which is a full-process data processing technology system based on data set delivery of labeled data. Focus on enhanced data calculation methods and production data business capabilities. In order to improve the efficiency of query and data calculation, and then improve and innovate applications, lay the foundation for subsequent big data processing and applications.

As the user's business needs change, the system will automatically recommend a suitable data label for the user, and automatically identify the logic between the physical structure in the data table and the historical data. Utilizing the powerful scalability, fast search efficiency and the powerful functions of highly accurate mark-based analysis technology, the search speed of massive telemetry data is significantly enhanced. Based on the tag center, an integrated logic model is created on the big data resources. The "tag" model log view can provide a variety of data service modules for users in a variety of different business scenarios, including drawing image analysis, rule prompts, text mining, personalized recommendations, relational networks, etc. The combination of interfaces can realize rapid analysis and application program construction. Interactive data metadata is created through a set of tags, and the data structure is centrally managed in the system. The system automatically creates a data source list while performing interactive data source security management and control.

3.5 Data Management

Data management includes a data management mechanism, which consists of three parts: technology, tools, and systems. It usually supports data governance to ensure efficient and effective work related to data governance.

3.6 Construction of Information Processing Cluster Framework

Hadoop is based on the Java language, so it has strong cross-platform capabilities. It can be well adapted to the system environment of Windows, Linux and Mac systems.

(1) Hardware description

The Hadoop platform uses four computers, one of which is the master node and the other three are child nodes. The main computer parameters are: CPU: dual-core processor, 2.40 GHz, memory: 8.0 GB, system type: 64-bit operating system, hard disk: 2 T.

(2) Software description

1) Linux system selection

Linux systems can use virtual machines or dual-system installations together with Windows, so they will not affect other computer application requirements during the testing phase. Therefore, this experiment uses the Linux Ubuntu operating system for corresponding experimental operations. For this experiment, choose Ubuntu LTS 15.04.

2) Hadoop version selection

This article uses Hadoop 3.6.0 (stable) version.

4 Test of the Large-Capacity Data Processing System of the Main Distribution Network Based on the Information Processing Cluster Framework

4.1 Algorithm and Accuracy of the System

In this experiment, in order to test the accuracy of the algorithm and the system, the test process is as follows: (1) the collected data is divided into five data subsets, which is convenient for comparison and verification of the experiment. (2) A total of 3 experiments were carried out. In each experiment, three sets of data subsets are selected as the training set, and one set is used as the validation set. The experimental results are shown in Table 1.

Table 1. Algorithm and accuracy of the system

	1	2	3
Subset 1	67%	68%	71%
Subset 2	71%	70%	72%
Subset 3	70%	71%	69%
Subset 4	73%	72%	70%
Subset 5	72%	71%	72%

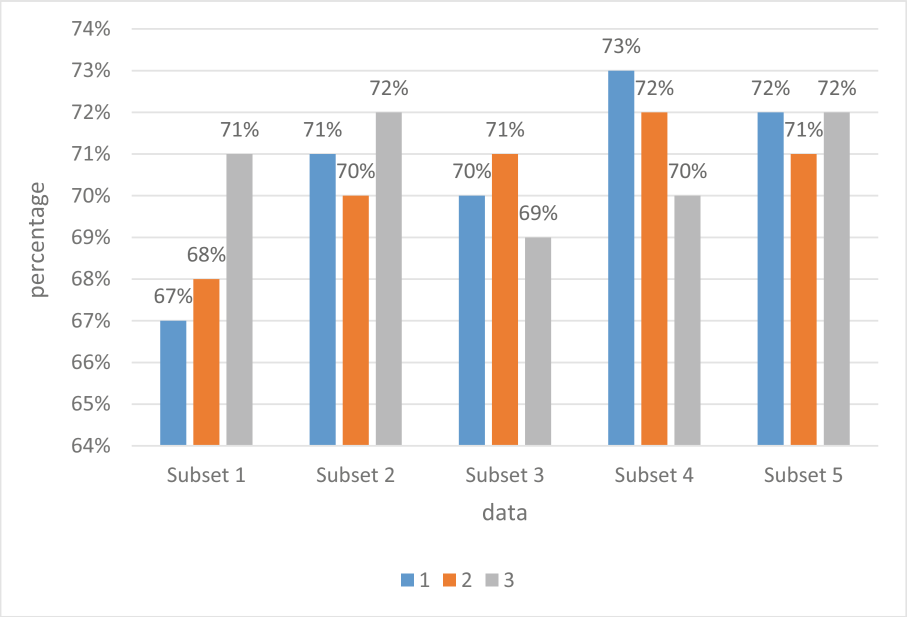

Fig. 2. Algorithm and accuracy of the system

It can be seen from Fig. 2 that the data processing accuracy rate of the system is about 70%. It can be seen that the accuracy of the parallel algorithm implemented by the system is within an acceptable range.

4.2 Validity Test

In this experiment, in order to test the efficiency of the system and algorithm, a total of three data sets are created, and the serial algorithm (the node is set to 1, which is equivalent to a serial function) is compared with the parallel algorithm. The experiment is based on the use of Data for predicting the wrong category. The five data sets are (10 million, 20 million, 30 million, 40 million, and 50 million data volumes). The experimental results are shown in Table 2.

Table 2. System effectiveness test

	Serial time (ms)	Parallel time (ms)
1	150006	224571
2	191220	306576
3	233453	343255
4	312452	421341
5	432145	456325

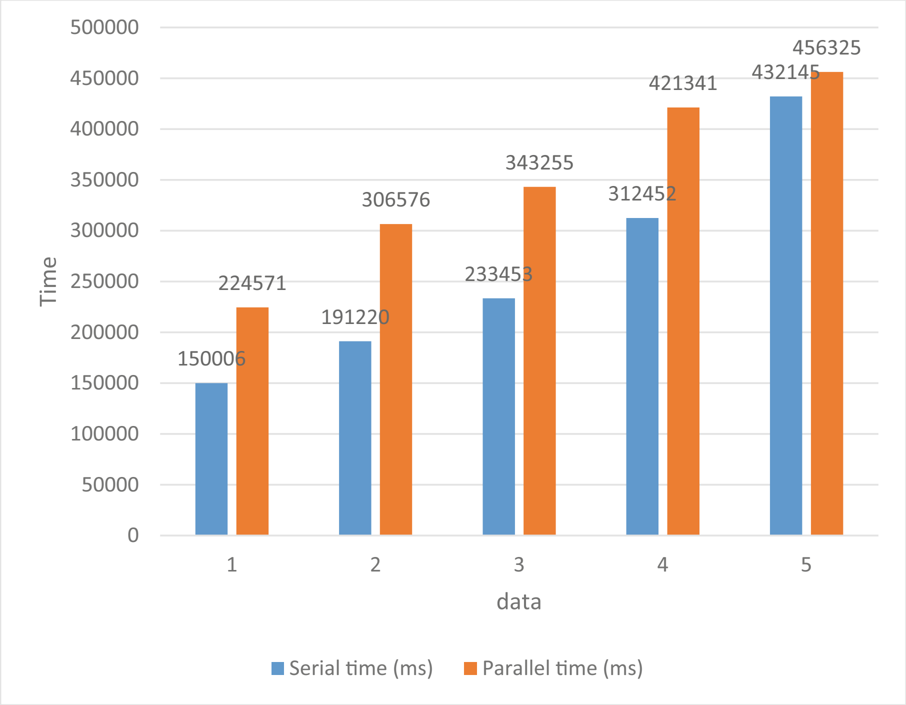

Fig. 3. System effectiveness test

It can be seen from Fig. 3 that compared with serial time, parallel test time is significantly reduced. This experiment shows that the parallel algorithm based on Hadoop has been improved more successfully, and the parallel processing performance of the platform is better.

5 Conclusions

This paper studies the large-capacity data processing of the main distribution network based on the information processing cluster framework. After analyzing the characteristics of the main power grid data, according to the situation, a large-capacity data processing system for the main distribution network based on the information processing cluster framework is designed. Then the designed system is tested. According to the accuracy test of the system, the test result shows that the data processing accuracy of the system is about 70%, which basically meets the system requirements.

Acknowledgements. "(Fund Project: China Southern Power Grid Corporation Key Technology Project Funding (R&D and Engineering Demonstration of Edge Computing Cluster Technology for Provincial Dispatching Main Station of Smart Grid, GXKJXM20190619)".

References

1. Wang, Z., Ng, D., Wong, V., et al.: Robust beamforming design in C-RAN with sigmoidal utility and capacity-limited Backhaul. IEEE Trans. Wirel. Commun. **16**(9), 5583–5598 (2017)
2. Silva, D.A.N.S., Souza, L.C., Motta, G.H.M.B.: An instance selection method for large datasets based on Markov Geometric Diffusion. Data Knowl. Eng. **101**(Jan.), 24–41 (2016)
3. Ferreira, R.S., Bentes, C., Costa, G., et al.: A set of methods to support object-based distributed analysis of large volumes of earth observation data. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. **10**(2), 681–690 (2017)
4. Saleem, A., Khan, A., Malik, S., et al.: FESDA: fog-enabled secure data aggregation in smart grid IoT network. IEEE Internet Things J. **7**(7), 6132–6142 (2020)
5. Lowik, S., Kraaijenbrink, J., Groen, A.: The team absorptive capacity triad: a configurational study of individual, enabling, and motivating factors. J. Knowl. Manag. **20**(5), 1083–1103 (2016)
6. Yan, C.G., Wang, X.D., Zuo, X.N., et al.: DPABI: data processing & analysis for (resting-state) brain imaging. Neuroinformatics **14**(3), 339–351 (2016)
7. Jindal, A., Dua, A., Kaur, K., et al.: Decision tree and SVM-based data analytics for theft detection in smart grid. IEEE Trans. Industr. Inf. **12**(3), 1005–1016 (2016)
8. Hsieh, K., Ebrahimi, E., Kim, G., et al.: Transparent offloading and mapping (TOM): enabling programmer-transparent near-data processing in GPU systems. Comput. Archit. News **44**(3), 204–216 (2016)
9. Mehenni, A., Alimazighi, Z., Bouktir, T., Ahmed-Nacer, M.: An optimal big data processing for smart grid based on hybrid MDM/R architecture to strengthening RE integration and EE in datacenter. J. Ambient Intell. Humaniz. Comput. **10**(9), 3709–3722 (2018). https://doi.org/10.1007/s12652-018-1097-4
10. Vanfretti, L., Olsen, S.H., Arava, V., et al.: An open data repository and a data processing software toolset of an equivalent Nordic grid model matched to historical electricity market data. Data Brief **11**(C), 349–357 (2017)
11. Al-Rubaye, S., Kadhum, E., Ni, Q., et al.: Industrial internet of things driven by SDN platform for smart grid resiliency. IEEE Internet Things J. **6**(1), 267–277 (2019)
12. Hu, J., Yang, K., et al.: Guest editorial: smart grid inspired data sensing, processing and networking technologies. Mob. Netw. Appl. **24**(5), 1699–1700 (2019)