Application Research of Big Data Real-time Processing Technology in Smart Grid

Sheng Zhou¹, Hongfa Li¹, Jianglong Zhang¹, Shuling Wu¹, Shiton Chen², Xiaohua Wu² and Tao Jin¹,*

¹State Grid Fujian Electric Power Co., Ltd, Fuzhou 350003, China
²Anhui Jiyuan Software Co., Ltd. Hefei 230088, China

*Corresponding author email: jintly@fzu.edu.cn

Abstract. Big data is valuable, and big data processing technology can efficiently process massive data and get the value in big data, so the research and application of big data technology become popular nowadays. Smart grid is one of the most potential field for application of big data technology, applying big data technology to smart grid can improve the reliability and economy of smart grid. The operating conditions of power system changes rapidly, therefore, real-time processing technology became the focus of application field. This paper reviewed the data sources, data characteristics and data application scenarios of big data in smart grid, analyzed data processing modes of big data processing technology: batch processing and stream processing, and compared the realization methods of real-time big data processing, summarized the application of big data real-time processing technology in smart grid, and the challenges in the application, and finally, prospected the application of real-time big data processing technology in smart grid.

Keywords: Big Data; Smart Grid; Real-Time Processing; Power system.

1. Introduction

With the continuous development of Internet information technology, the amount of data generated by human beings on the Internet is also increasing and accumulating with an explosive speed. In 2018, the world can generate 2.5 exabytes of data every day[1]. These huge amounts of data have an important relations to the Internet, finance, industry, transportation and other fields[2]. We are in the "big data era". There is no standard academic definition of big data, but in terms of characteristics, big data has 4V characteristics: volume (large data volume), velocity (high speed of data generation), variety (diversity of data sources and data structures), and value (the total value of data is high but the density is low)[3]. Big data technology mainly analyzes and summarizes massive data through statistical search, analysis, comparison, and clustering, and finds out the implicit interrelationships in the data set [4], so as to dig out the required information. Big data has also given birth to the "fourth paradigm" of data-intensive scientific research [5]. This new scientific research model provides new methods and ideas for solving many current problems. Today, big data and related technologies have been widely recognized and applied.

Power grid is very important areas for big data technology application. Firstly, with the continuous advancement of smart grid construction, the number of monitoring devices in the grid is increasing rapidly, and the data generated in the process of power generation, transmission, transformation, distribution, and power enterprise marketing increases exponentially; on the other hand, when the smart grid is running, it is necessary to monitor the state of the power grid in real-time, and perform comprehensive analysis and processing on the data in the power grid to achieve optimal management of
the entire power system. These conditions and requirements of the smart grid provide the foundation and driving force for the application of big data technology in the smart grid. At present, there are many big data technology application schemes. The literature [6, 7] applies big data processing technology to data analysis and load forecasting of smart grid, so that the operation of the grid can be more intelligent. Research on the real-time processing technology of smart grid big data is significant to the improvement of the smart grid real-time monitoring, online real-time fault diagnosis, and the construction of an economical and reliable smart grid. In this article the sources, characteristics and application scenarios of big data in smart grids was described; the modes of big data processing technology and the realization methods of big data real-time processing technology was introduced; application of big data real-time processing technology in smart grid was reviewed; the challenges in application were summarized; based on the above, the application of big data real-time processing technology in smart grid was prospected.

2. Big Data Analysis in Smart Grid

The smart grid connects various sensors and equipment through the network, so as to summarize, integrate, and analyze the various information collected in the grid, and accordingly reduce the operating cost, improve efficiency and reliability, and make the operation and management of the grid system optimized [6]. In this process, a large number of smart meters, smart equipment monitoring systems, grid management systems, and many other data sources related to the grid are constantly generating data. These massive amounts of data from all parts of the smart grid together forms the smart grid big data. The source and types of data was summarized in the following table:

| Data source            | Data type                                      |
|------------------------|------------------------------------------------|
| Grid internal data     | Voltage, Current, Phase, Frequency, etc.       |
| Grid enterprise data   | Enterprise management data, Electricity consumption data, Marketing data, etc. |
| Grid external data     | Economic population data, Meteorological data, Geographic data, etc. |

Smart grid has abundant sources and large data volume, which can support many applications. The focus is mainly on three aspects: one is to support grid operation and operation; the second is to serve the society and support the development of other industries; and the third is to serve power users. These three aspects include many technical fields, summarized in the Table 2. The table showed that the current smart grid big data application fields can be roughly divided into two categories: One category is applications that require deep data mining and do not require high data processing speed, such as grid planning, user energy efficiency analysis, and marketing business assistance Analysis, etc.; the second category is applications that only need to perform simple analysis and processing of data but require higher data processing, such as online monitoring and fault diagnosis of power grid equipment, power system transient stability analysis and control.
Table 2. Application direction and field of big data in smart grid[7].

| Direction                  | Field                                                                 |
|----------------------------|----------------------------------------------------------------------|
| Support grid operation     | Power system transient stability analysis and control,                |
|                            | Power grid equipment online monitoring and fault diagnosis,           |
|                            | Load Forecasting,                                                    |
|                            | Wind turbine, photovoltaic power generation forecast,                |
|                            | Distribution network fault location,                                 |
|                            | Grid equipment asset management,                                    |
|                            | Power grid planning, etc.                                            |
| Serve the society, support| Socioeconomic situation analysis and forecast,                        |
| the development of other   | Related policy formulation and effect analysis, etc.                  |
| industries                 |                                                                      |
| Service for power users    | Power supply and demand management,                                  |
|                            | User energy efficiency analysis,                                     |
|                            | Analysis and optimization of power supply services,                  |
|                            | Marketing business auxiliary analysis, etc.                           |

3. Big Data Real-time Processing Technology

The traditional data processing mode is to gather data into computing center for calculation and processing. The 4V characteristics of big data make the big data processing system need to perform a large number of calculations in a short time. Faced with massive data, traditional data processing method, can no longer meet the demand, due to the limitation of data storage capacity and computing efficiency. The current major big data processing technology is distributed computing. The principle of distributed computing is shown in Figure 1. Distributed computing technology can process big data, but the implementation schemes of different distributed computing technologies have different processing modes of computing tasks, which also makes different processing modes have different characteristics.

Big data processing modes can be divided into two [8]: batch processing and stream processing. The real-time processing technology of big data can be implemented in two modes, batch processing and stream processing as well.

![Figure 1. Principles of Distributed Computing.](image)

3.1. Real-time Processing Technology Based on Batch Processing

Batch processing means that when the computing system processes computing tasks, it divides the tasks into multiple fixed-size batches for processing. When a batch is processed, the computing system no longer accepts data, and the next batch is processed after one task.

MapReduce [9] is a typical distributed batch processing architecture. It was first developed by Google to solve search-related problems. It can handle large-scale data, and its structure is simple and has good versatility. Therefore, the model has been received great attention. At present, Hadoop, a big data processing tool based on MapReduce, has been widely used in big data research field. The MapReduce architecture is mainly composed of Master and Worker nodes. Its data processing function is
implemented by two operations, Map and Reduce. The data passed in the process are all key-value pairs <key, value>. Its operating mode is shown in Figure 2.

It can be seen from the batch processing mode that when this data processing mode processes multiple batches of data, there must be a delay between batches. If the processing delay is high, the real-time processing needs cannot be met. Therefore, to use the batch processing mode to realize the real-time processing of big data, it is necessary to reduce the processing delay between multiple batches of data. Many developers have conducted research and exploration on this.

3.2. Real-time Processing Technology Based on Stream Processing

Stream processing means that when a computing system is processing a computing task, once the task is started, each computing node in the system will continuously perform data input, processing, and data output operations according to the established arrangement, just like the assembly line in the workshop, unless the operator terminates the task, and the task will go on forever. Stream processing has been widely used, and currently there are stream processing systems such as Smaza, and Storm. Storm [10] is a widely used stream processing architecture, its structure is similar to MapReduce, but the biggest difference from MapReduce is that Storm uses a Topology composed of Worker nodes named Spout and Bolt to process computing tasks, and can achieve task flow deal with. The operation mode is shown in the Figure 3.

When performing calculation tasks, the Spout node in the Topology continuously reads data from the data source and distributes the read data in a sequence of Tuple tuples. The Bolt node receives the Tuple transmitted by the Spout node, and completes the processing of the custom function. Then use Tuple to continue transmitting to the next Bolt for further processing or output the result directly. The stream processing mode represented by Storm processes data in a task topology, and the task will never end after it starts, while the batch processing mode represented by MapReduce ends the task after the current data is processed, and it will be required when the next task comes. Restart and initialize the system, so when processing continuous tasks, stream processing mode has lower data processing latency than batch processing mode. The delay of stream processing is low, but the data throughput is small, so when using...
stream processing, it is necessary to reasonably perform task decomposition and design topology, which is also a difficult point in the application of stream processing technology.

3.3. Real-time Processing Technology Based on Hybrid Processing

Hybrid processing is a data processing method that has capabilities for stream processing and batch processing. Realizing stream processing through batch processing framework and realizing batch processing through stream processing framework are two ways of hybrid processing. Spark Stream [11] is a distributed platform that uses batch processing mode to implement data processing as Figure 4. Unlike MapReduce, in order to further reduce the data processing delay, Spark Stream saves intermediate files in memory, shortening the data reading time and reducing processing time required for each batch. Spark Stream divides the data into a large number of micro-batches according to the time unit and converts them into RDD (Resilient Distributed Dataset) containing data in a specific time interval for batch processing. Continuous fast batch processing in a short time unit makes RDD form DStream data Flow, its mode is as follows:

![Figure 4. Spark Stream data flow.](image)

The two DStreams in the above figure indicate that Spark Stream can not only generate data streams in data sources, but also use existing RDD data to generate new data streams, and then process them according to different purpose. The Spark Stream model enables it to have both batch processing and quasi-stream processing capabilities. Flink [12] has the ability of stream processing framework to process batch tasks. Flink has a stream processing framework, which can read continuous data streams from beginning to end similar to Storm, or read batches of data streams with a set time window similar to Spark Stream for batch processing. Hybrid processing is a general solution for data processing and the best choice for diversified big data processing tasks. However, it has advantages and disadvantages compared with simple batch processing and stream processing in terms of implementation cost, data throughput, and processing delay. A simple comparison is as follows:

| Processing mode | Data throughput | Delay | Realization cost |
|-----------------|-----------------|-------|------------------|
| Batch           | high            | Minutes | low              |
| Stream          | low             | Milliseconds | high            |
| Hybrid          | adjustable      | Sub seconds | high            |

4. Application of Big Data Real-time Processing Technology in Smart Grid

The potential application areas of big data in smart grids are very rich. In recent years, big-data-related technologies have become a research hotspot. Big-data-related technologies have been developed by leaps and bounds. The big data real-time processing system can meet real-time requirements, and has many applications in smart grid status monitoring, fault diagnosis, and power system transient stability analysis.

The state monitoring of smart grid is divided into state monitoring and state control of smart grid as Figure 5. The status monitoring of the smart grid is to obtain the running status of the system through the data measured by various measuring devices in the smart grid. The system continuously measures and feedbacks the result at a certain frequency is the monitoring process. Smart grid monitoring is the
basis of system regulation and a prerequisite for grid fault diagnosis. Only accurate monitoring can obtain the correct system status and proceed with the next regulation or diagnosis.

The challenge of power grid operation monitoring is that the data monitored by the system is multi-source, heterogeneous and huge. Compared with traditional operation monitoring technology, the advantage of big data real-time processing technology is that it can process multi-source heterogeneous massive operation data in real time. The operating status of the system can be obtained, and predicted based on the analysis of the massive data in the power grid, so that the dispatching of the power grid is more reasonable.

For the stable and economic operation of the power grid, it is needed to regulate the active and reactive power during oftenly, and the working conditions of the system change very rapidly, and it is difficult for traditional regulation methods to achieve real-time response. In contrast, big data real-time processing technology can monitor the grid status in real time, and can predict the load based on historical data. According to the real-time status and forecast data of the grid, the grid can be finely regulated, thereby improving the reliability and economy of the smart grid.

When the smart grid system detects a fault, it needs to diagnose and locate the fault. The faults in the smart grid can be divided into component faults, equipment faults and system faults. Fault diagnosis has a direct impact on the elimination of faults, and timely and accurate fault diagnosis The self-healing ability and reliability of the power grid are of great significance.

In operation, big data real-time processing technology can also be combined with artificial intelligence and machine learning technology to form an AI online fault diagnosis and analysis system as Figure 6, which can realize the functions of power grid fault prediction, online diagnosis and self-healing.

Big data real-time processing technology is very suitable for smart grid, but most current application researchs of big data real-time processing technology in smart grid is still in the experimental exploration stage. Figure 7 is a practical application in power system, the average processing time after fusion is between 90-300 ms, which meets the time requirements of big data processing.
5. Conclusion
At present, the application of smart grid big data real-time processing technology is still in the research trial stage of scientific researchers or a few companies. The promotion of big data real-time processing technology in smart grids is facing many challenges. How to combine big data technology with IoT technology and artificial intelligence technology in power system, it is an important work to meet the needs of smart grid in the future.

Acknowledgement
This paper is supported by Science and Technology Project of State Grid Corporation of China (Research on Key Technology of Health Evaluation and Real-time Early Warning of Information System for Massive Logs, No.52130M190008).

References
[1] M. S. Hadi, A. Q. Lawey, T. E. H. El-Gorashi, and J. M. H. Elmirghani, "Big data analytics for wireless and wired network design: A survey," Comput. Netw., vol. 132, pp. 180-199, 2018.
[2] D. Wang, Y. Song, and Y. Zhu, "Information Platform of Smart Grid Based on Cloud Computing," Automation of Electric Power Systems, vol. 34, no. 22, pp. 7-12, 2010.
[3] X. Meng and X. Ci, "Big Data Management: Concepts, Techniques and Challenges," Journal of Computer Research and Development, vol. 50, no. 1, pp. 146-169, 2013.
[4] K. M. Tolle, D. S. W. Tansley, and A. J. G. Hey, "The Fourth Paradigm: Data-Intensive Scientific Discovery," Proc. IEEE, vol. 99, no. 8, pp. 1334-1337, 2011.
[5] P. D. Diamantoulakis, V. M. Kapinas, and G. K. Karagiannidis, "Big Data Analytics for Dynamic Energy Management in Smart Grids," Big Data Research, vol. 2, no. 3, pp. 94-101, 2015.
[6] R. Shyam, B. Ganesh H.B, S. Kumar S, P. Poornachandran, and K. P. Soman, "Apache Spark a Big Data Analytics Platform for Smart Grid," Procedia Technology, vol. 21, pp. 171-178, 2015.

[7] D. Zhang, X. Miao, L. Liu, Y. Zhang, and K. Liu, "Research on Development Strategy for Smart Grid Big Data," Proceedings of the Chinese Society of Electrical Engineering, vol. 35, no. 1, pp. 2-12, 2015.

[8] X. Cheng, X. Jin, Y. Wang, J. Guo, T. Zhang, and G. Li, "Survey on Big Data System and Analytic Technology," Journal of Software, vol. 25, no. 9, pp. 1889-1908, 2014.

[9] J. Dean and S. Ghemawat, "Mapreduce: Simplified data processing on large clusters," Commun. ACM, vol. 51, no. 1, pp. 107-113, 2008.

[10] A. Toshniwal et al., "Storm@twitter," presented at the Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data, Snowbird, Utah, USA, 2014.

[11] M. Zaharia, T. Das, H. Li, T. Hunter, S. Shenker, and I. Stoica, "Discretized streams: fault-tolerant streaming computation at scale," presented at the Proceedings of the Twenty-Fourth ACM Symposium on Operating Systems Principles, Farmington, Pennsylvania, 2013.

[12] P. Mika, "Flink: Semantic Web technology for the extraction and analysis of social networks," J. Web Semant., vol. 3, no. 2-3, pp. 211-223, 2005.