IOP Conf. Series: Materials Science and Engineering **394** (2018) 042116 doi:10.1088/1757-899X/394/4/042116

Fault Diagnosis and Trace Method of Power System Based on Big Data Platform

Lei Wang*, Lingling Shang, Mengchao Ma and Zhiguang Ma

Power Grid Maintenance Training Department, State Grid of China Technology College, Jinan 250002, China

*Corresponding author e-mail: 37601391@qq.com

Abstract. This paper put forward a new power grid fault trace method based on big data platform. It extend the data source to transformer substation and deal with the fault date by the technology of Spark. This method can solve the problem of mass data processing. This paper analyses fault information by data mining and trace the malfunction of protection or circuit breaker by decision tree. It can give the reason of fault and optimize the function of fault diagnosis system. Compared with the traditional fault diagnosis system, which based on alarm messages, this method can use every monitoring data in substation and give the reason of fault.

1. Introduction

The fault diagnosis of power system is a basic subject for the realization of the self-healing function of smart grid. The research results in recent years are very rich, and some applications have been obtained in the production field. The current fault diagnosis technology mainly relies on protection starting, protection action information and circuit breaker tripping signal for fault reasoning. For example, analytic model for fault diagnosis in power systems considering malfunctions of protective relays and circuit breakers [1], power grid fault diagnosis based on reasoning chain [2], fuzzy cellular fault diagnosis based on radial basis function neural network [3] and petri net [4]. The above fault diagnosis algorithm needs timely and stable real-time data to support. However, due to the lack of effective data preprocessing mechanism and mutual independence among various information systems, the following two disadvantages are caused. First, the alarm information of substations at all levels at the same time will be collected from the dispatching center, which can easily cause data congestion or even distortion. Second, the substation layer before failure warning information, such as optical fiber sampling abnormality, GOOSE link chain scission, grade an equipment online monitoring and forecast information in each substation server storage alone, serious information barriers.

On the other hand, the rise and improvement of big data technology has caused the fundamental change of scientific research pattern. Big data technology is based on a large number of source, type and complex high speed to capture the data, discovery and analysis, extract data using the method of economic value of the technical system or technical architecture. The characteristics of large data technology scale, diversity and high speed can make up the disadvantage of power failure information system.

At present, the intelligent station is equipped with a variety of information acquisition system, which is designed to effectively monitor all kinds of equipment in the station. Numerous information

brings the huge amounts of data at the same time, also brings great inconvenience to maintenance personnel, in the face of two seconds of time will turn screen in real time the alarm monitor, inevitably there are omissions, and this reduces the utilization of information resources to a certain extent.

In this paper, the starting point is using the results of the fault diagnosis after the failure to protect or circuit breaker disoperation situation is analyzed, and the use of big data architecture to improve the power system fault information collection way, focus on the use of the data mining technology to protect or circuit breaker refusing action of reverse trace analysis. The function of fault diagnosis algorithm is greatly improved by utilizing all kinds of information sources in power system.

2. Power Grid Fault Information Architecture Based On Big Data Platform

Compared to the traditional database cluster technology, large data with high fault tolerance, high scalability, data diversity, the characteristics of the processing accuracy, its processing mode from the traditional database cluster evolution to graphs and HDFS massively parallel processing architecture, realizes the task decomposition and merge the results, so as to realize the unlimited extension can process the data scale [5].

Big data platform to use its resources optimization scheduling, high-speed data transmission mechanism, a copy of the data management technology of database cluster out of rush-hour network data transmission bottleneck and error problem.

Given fault diagnostic procedures need to deal with the in a short period of time after a lot of electric parameters information or the electric parameters, so the project on the basis of previous work, choose a big data platform as the data support environment of fault diagnosis system, using distributed storage technology to get the failure data and management of substation layer, it can avoid data in scheduling the excessive congestion, to the diagnosis of upper applications provide a more stable and fast data interface.

2.1. Concepts of big data and implementation methods.

At present, both traditional and intelligent power stations are equipped with many monitoring systems. According to the 8 real-time alarm information per second, about 700 thousand days per day, the storage period is one month, and the rolling alarm information in real-time database is about 20 million. For the entire power grid equipment monitoring platform, to store monitoring or management of data more large, rely on the traditional relational database is difficult to meet the demand for smart grid fault diagnosis and self-healing, this article attempts to big data technology to improve data grid environment now, an alternative architecture and gives large fault the data platform

Since large data is the theory that has risen in recent years, its connotation and extension are constantly changing, and technical standards are also relatively open source. Big data is data sets that are so voluminous and complex that traditional data processing application software are inadequate to deal with them. Big data technology is obvious for the smart grid currently under construction: big data will provide reliable data sources and stable data quality for smart grid, and provide efficient management mechanism for all dispatching centers to control the state of power grid. SCADA, GOOSE and EMS can be used for fault diagnosis of data sources through the corresponding data engine registered as part of the smart grid big data platform. Through Hadoop YARN on the data source of job scheduling and unified data management, which maintain independent operation of the data acquisition system, capable of parallel processing of power system fault keep high throughput data access control.

The most popular open source software in the field of big data is Apache Hadoop [6]. Hadoop is an open source software framework that enables distributed processing of large amounts of data. Users can build distributed computing platforms for applications based on the framework they provide. According to the Hadoop 2.0 protocol framework, the architecture of the power grid fault data platform proposed in this paper is shown in Fig.1.

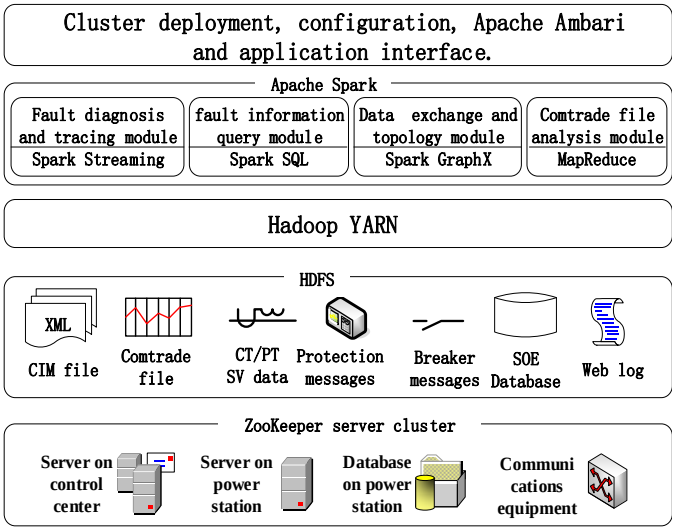

Figure 1. Big data platform of power grid fault

In the failure big data platform architecture shown above, Zookeeper server cluster is bottom layer. Zookeeper is a centralized service for maintaining configuration information, naming, providing distributed synchronization, and providing group services. All of these kinds of services are used in some form or another by distributed applications. Each time they are implemented there is a lot of work that goes into fixing the bugs and race conditions that are inevitable. Because of the difficulty of implementing these kinds of services, applications initially usually skimp on them, which make them brittle in the presence of change and difficult to manage. Even when done correctly, different implementations of these services lead to management complexity when the applications are deployed. In this architecture, Zookeeper control all the servers in power grid stations. For example the server on control center, the server on power station and database on power station. Because the communication equipment can produce and send messages, they are in the control of Zookeeper.

The second layer is HDFS. HDFS (Hadoop Distributed File System) is the primary distributed storage used by Hadoop applications. A HDFS cluster primarily consists of a NameNode that manages the file system metadata and DataNodes that store the actual data. The HDFS Architecture Guide describes HDFS in detail. This user guide primarily deals with the interaction of users and administrators with HDFS clusters. The HDFS architecture diagram depicts basic interactions among NameNode, the DataNodes, and the clients. In the malfunctioning data platform architecture, HDFS is used to manage all the data available for fault diagnosis. For example, the CIM file, COMTRADE file, CT/PT data, alarm messages of protection and breakers, SOE and Web log. Compared with traditional distributed database, HDFS has better coordination ability of distributed data, and can increase the fault tolerance of the system through backup storage technology.

Hadoop provides resource scheduling and management for all kinds of application computing through YARN. YARN (Yet Another Resource Negotiator) supports the notion of resource reservation via the Reservation System, a component that allows users to specify a profile of resources over-time and temporal constraints, and reserve resources to ensure the predictable execution of important jobs. The Reservation System tracks resources over-time, performs admission control for reservations, and dynamically instruct the underlying scheduler to ensure that the reservation is fulfilled. The fundamental idea of YARN is to split up the functionalities of resource management and job scheduling/monitoring into separate daemons. The role of YARN in the framework is to provide a unified resource scheduling service for multiple subprograms in the fault diagnosis and to share the cluster resources. The subprograms includes fault diagnosis program, power gird topology program, and COMTRADE file analysis program and so on. YARN separations the data resources from the

diagnostic program in the framework, and the advantage is that when the upper level of the diagnostic algorithm changes, it does not affect the HDFS.

The upper layer of YARN is a fault diagnosis program with Apache Spark as the main computing program. Apache Spark is an open-source cluster-computing framework [7]. Originally developed at the University of California, Berkeley's AMP Lab, the Spark codebase was later donated to the Apache Software Foundation, which has maintained it since. Spark provides an interface for programming entire clusters with implicit data parallelism and fault tolerance. Apache spark is compatible with the apache HADOOP API, so it can read and write the resources in HDFS, and it can seamlessly connect to the YARN. Compared with MapReduce, the Spark completely based on memory computing, so the operation performance is more superior and the handling of the real-time processing system is more efficient.

2.2. Four Modules of Fault Diagnosis and Tracing

In view of the above analysis, this paper divides the post-failure processing program into four modules. The first one is fault diagnosis and tracing module. This module is the main part of the diagnostic program. It is responsible for analyzing the alarm data and finding out the reason of the failure component or the protection and the circuit breaker. Because the processing priority of this module is the highest, this article selects the Spark Streaming component to process the module. Spark Streaming brings Apache Spark's language-integrated API to stream processing, letting you write streaming jobs the same way you write batch jobs. Spark Streaming component is characterized by the ability to receive real-time data, which can be divided into multiple batches for processing by cycle, and the processing cycle is short, usually in milliseconds. Spark flow calculation program can maximum fit the characteristics of the fault diagnosis process real-time processing, so this paper use it to handle fault diagnosis and tracing program.

The second one is fault information query module. At present, in the hierarchical dispatching mechanism of power grid stipulates, only the alarm information of accident level can upload the dispatching master station. So many useful early warning information for fault diagnosis is only stored in the server of substation. This article selects Spark SQL to query the data in the station to provide support for fault tracing. Spark SQL is a Spark module for structured data processing. Unlike the basic Spark RDD API, the interfaces provided by Spark SQL provide Spark with more information about the structure of both the data and the computation performed. Spark SQL can query structured data inside Spark programs, using either SQL or a familiar DataFrame API. It enables structured queries against distributed data sets and has highly performant. So this paper use Spark SQL to search alarm messages in power station.

The third one is data exchange and topology module. The fault diagnosis needs the data exchange between the main station and the substation and the topology analysis of the power grid. This paper use GraphX as the tool to handle topology data. GraphX is a new component in Spark for graphs and graph-parallel computation. At a high level, GraphX extends the Spark RDD by introducing a new Graph abstraction: a directed multigraph with properties attached to each vertex and edge. To support graph computation, GraphX exposes a set of fundamental operators as well as an optimized variant of the Pregel API. In addition, GraphX includes a growing collection of graph algorithms and builders to simplify graph analytics tasks. It can complete the data model exchange between the IEC61850 power station system and the IEC61970 in control center. In addition, it can be able to make use of SVG graphics files generated by sub stations to form a full grid topology. Next, it can find out subnet without generator then the scope of the fault diagnosis will be determined.

The forth one is COMTRADE file analysis module. As an analysis part of the failure, the module is designed as a batch processing system. The processing object of this module is dozens of megabytes of recording file, so it is not suitable for memory operation with Spark. In this paper, Hadoop distributed computing component MapReduce is selected for offline computing. MapReduce is a programming model and an associated implementation for processing and generating big data sets with a parallel, distributed algorithm on a cluster. The MapReduce orchestrates the processing by

marshalling the distributed servers, running the various tasks in parallel, managing all communications and data transfers between the various parts of the system, and providing for redundancy and fault tolerance. Compared to Spark, MapReduce writes the exchange data to the external disk file system, so the computing speed is lower than Spark. But the advantage of MapReduce lies in large block data processing, and the data proximity replication strategy also makes it good for fault tolerance. It is suitable for the analysis and storage of recorded wave files. The size of COMTRADE file is 30M to 100M, so MapReduce is adopted in this module.

The top of the framework is the cluster deployment, configuration, management interface, and application interface layer. This interface is implemented by Apache Ambari. The Apache Ambari is aimed at making Hadoop management simpler by developing software for provisioning, managing, and monitoring Apache Hadoop clusters. Ambari provides an intuitive, easy-to-use Hadoop management web UI backed by its APIs. The application interface can provide diagnostic results for the self-healing process of the smart grid, or provide a fault summary at the scheduling center.

3. Fault Diagnosis and Tracing Application Design Based On Big Data Platform

In the big data platform, there are three data processing system. They are real time processing system, quasi-real time processing system and Non-real-time processing system. In the real-time processing system, the accident level information is composed of protection start, protection action and circuit breaker tripping information. The power station will report the incident level information in real time, and the real time processing system will get highest level of response. Fault diagnosis algorithms based on accident level information, so it has the highest priority and the fastest response.

If the conclusion of the fault diagnosis algorithm is that both the protection and the circuit breaker are correct, the diagnostic brief report will be sent to the dispatcher or to the smart grid self-healing process. If some misoperations of breaker or protection are detected by fault diagnosis algorithm, the quasi-real time processing system will be on-line. The quasi-real time processing system will handle alarm messages of GOOSE, link break messages etc. However, some alarm messages, which are not important, will not be sent to control center, they only be stored in database of power station. This paper use data mining to design quasi-real time processing system [8].

Another data source is COMTRADE file. In view of the characteristics of large storage space and low demand for real time, a Non-real-time processing system is used in this paper. It provides specific fault phase, fault time and accurate fault location for fault diagnosis.

The big data platform use 4 steps to complete the fault diagnosis and tracing program.

Step 1. When the fault occurs, the alarm messages of accident level will be collected by control center. Then it triggers the fault diagnosis application. The application will submit a task application to the YARN cluster explorer. The resource manager is responsible for scheduling the running resources of each program module such as memory, CPU, network and disk space.

Step 2. The resource manager assigns Work Nodes (WN) to each task. The program code of fault diagnosis module, information query and other modules will be distributed to the corresponding WN. WN also provides feedback to the resource manager on the resource usage and program running status of the node.

Step 3. Each WN executes the program code. WN1 selects alarm information from HDFS and finds out fault section. If WN1 detect some misoperation of protection or breaker, WN2 will online. WN2 uses Spark SQL to query the station information from HDFS. It send the alarm messages in power station to WN1. WN3 extracts SVG graphic files conforming to the CIM model from HDFS. The power grid topology will be formed by the SVG graphic and passive region will be get. WN4 analyzes the COMRADE files generated after the failure.

Step 4. WN1 returns the fault diagnosis and tracking results to the fault diagnosis request interface. WN4 sends the offline analysis results such as fault phase and fault time to the diagnostic request interface.

The workflow of fault diagnosis and tracing procedure in shown in Fig. 2.

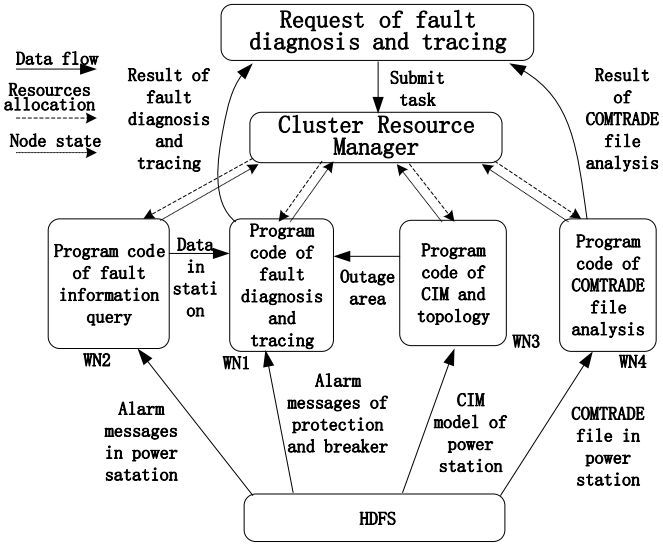

Figure 2. Flow chart of fault trace program

4. Application

The laboratory big data platform is configured with 5 physical workstations. The operating system is CentOS 6.4 and big data platform is Hadoop 2.6.0 with Spark-1.2.0. One of the workstations is used as the fault diagnosis server of the control center the other simulate power substation server.

In this paper, the information sampling statistics of three substations involved in the simulation experiment are analyzed. A permanent fault in the line is set between two substations of A and B. The protection of substation a refuse to work and the secondary backup protection of substation C cut off the fault current. There are about 70 million information in three substations. It contains 168 accident alarm information, 143 general alarm information and 122 warning alarm information. Take substation A as example, the information statistical table is shown in Table.1.

Table 1. Statistical table of information gain

Alarm Severity	Alarm information object.	Information increment
accident	protection action	9
accident	protection starting	64
accident	main transformer intelligent terminal starts	8
general	Reclosing action	4
general	GOOSE communication network broken	8
general	protection device merge unit network link error	8
general	Recorded Data accomplished	12
warning	Sampling anomaly of current and voltage	12
warning	Current and voltage sampling invalid	8
warning	Abnormity of current and voltage link	4
warning	Abnormity of fiber sampling in busbar measurement and control	4
warning	Other messages including Online monitoring and restored breaker etc.	12

Firstly, fault diagnosis algorithm is used to judge the fault section by accident alarm messages. The fault diagnosis program will find that the range of power outages has been expanded because malfunctions of protection on station A. Then the fault tracing algorithm will online. Secondly, we use decision tree [9] or k-means [10] algorithm of data mining to trace the reason of malfunction. We can get the decision that protection on station a failed to operate because of the breakdown of current and voltage sampling. The realization of whole model is shown in Fig.3.

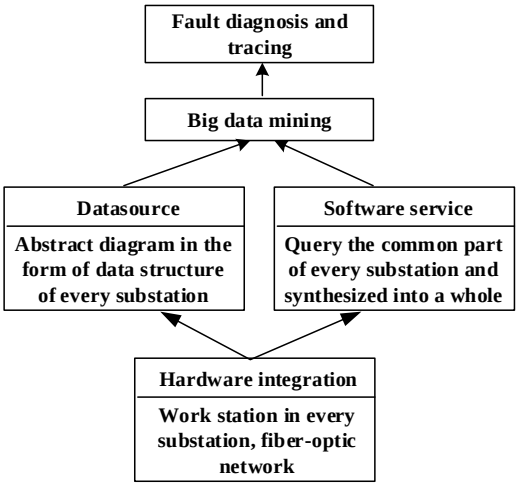

Figure 3. realization of the whole model

5. Conclusion

This paper establish a fault data collection model by taking advantages of the high efficiency and stability of big data on data collection with the purpose to provide fast, accurate and unified data for upper layer fault diagnosis program. At present, if switch is tripped by protection action, corresponding operator shall record accurate tripping time, details of protection action signal to be restored manually, optical signal and abnormities, and report such records to dispatcher on duty immediately. The process will last about 15 minutes. This paper use big data platform to collect alarm message and data mining to trace fault reason. This method will last about 1 minutes.

Acknowledgments

This work is supported by research and development program of higher education of Shandong province. (No. J17KB163).

References

[1] F. S. Wen and C. S. Chang, “A new approach to time constrained fault diagnosis using the Tabu search method,” J. Eng. Intell. Syst., vol. 10, no. 1, pp. 19-25, 2002.
[2] H. J. Lee, B. S. Ahn, and Y. M. Park, “A fault diagnosis expert system for distribution substations,” IEEE Trans. Power Del., vol. 15, no. 1, pp. 92-97, Jan. 2000.
[3] J. C. Tan et al., “Fuzzy expert system for on-line fault diagnosis on a transmission network,” in Proc. IEEE Power Engineering Soc. Winter Meeting, vol. 2, Jan. -Feb. 2001, pp. 775-780.
[4] L. Wang, Q. Chen, Z. Gao, et al., “Knowledge representation and general petri net models for power grid fault diagnosis,” IET Gener. Transm. Distrib, vol. 9, no. 9, pp. 866–873,2015.
[5] Big data [EB/OL]. [2012-10-02]. http://en. wikipedia.org/wiki/Big_data.
[6] Chuck Lam, James Warren. Hadoop in Action. Manning Publications. 2009.
[7] Spark: Cluster Computing with Working Sets. Matei Zaharia, Mosharaf Chowdhury, Michael J. Franklin, Scott Shenker, Ion Stoica. HotCloud 2010. June 2010.
[8] Wu Xingdong, Zhu Xingquan, Wu Gongqing, et al. Data Mining with Big Data [J]. IEEE Transactions on Knowledge & Data Engineering, 2014, 26 (1): 97-107.
[9] WZ Liu, AP White. The Importance of Attribute Selection Measures in Decision Tree [J]. Machine Learning, 1994, 15 (1): 25-41.
[10] C. Wang, J Bai, J Li, “Max-min K-means clustering algorithm and application in response signal feature extraction,” International Journal Of Applied Electromagnetic & Mechanics., vol. 39, no. 1, pp. 719–724, 2012.