An Approach to Smart Grid Online Data Mining Based On Cloud Computing

WANG Yuechao[1, a], CHEN Shihe[2]

1,2 Electric Power Research Institute
Duangdong Power Grid Co., Ltd, Guangzhou City 510080, China
[a]wangyuechao@163.com

***Abstract* — Smart Grid as a research focus and trends of power system is grid operation and management of the central nervous system, through various devices integrated service bus, integrated analysis of the information, improve grid efficiency and reliability. We consider the construction of smart grid background, combined with the actual situation and the needs of the power grid scheduling, proposed the use of data mining and integrated information technology platform to build applications systems programs to address the large number of scattered islands of information data and technical bottlenecks form makes it difficult to fast from data sources to extract valuable knowledge to effectively support questions grid scheduling decisions. Grid scheduling system construction business analysis to eliminate information silos, mining information from a large number of business data and extract knowledge, support intelligent scheduling business analysis and decision making.**

Keywords - Smart Grid; Data Mining; Cloud Computing; Map Reduce; Hadoop

I. INTRODUCTION

With the increasingly serious global energy issues, countries in the world to carry out research work smart grid. The ultimate goal is to build a smart grid electricity system covering the entire production process, including several aspects of power generation, transmission, substation, distribution, electricity and scheduling panoramic real-time systems. The support smart grid security, self-healing, basic green, strong and reliable operation of the grid is a panoramic real-time data collection, transmission and storage, as well as the cumulative mass rapid analysis of multi-source data [1-2]. As the amount of data thus deepening and advancing smart grid construction, power grid and equipment inspection / monitoring produced exponential growth, and gradually form a large circle of concern in today's information data, which requires appropriate storage and rapid processing technology as support. Due to widespread cloud application platform, it has accumulated a massive, multi-source heterogeneous data, which is in urgent need of people to study and theoretical analysis of this large data [3-5]. Currently, big data has become academia and industry research topics of common interest, in many areas received applications, has broad application prospects.

Online data mining is committed to analyze and understand the data, the data reveal knowledge hidden technology, which is digging out the implicit knowledge and information from a number of noisy data, and is currently one of the advanced data analysis tools. Online data mining models including classification model, clustering model, time series model, the associated mode, sequence mode. Grid data intelligent analysis system by means of data mining models in a variety of algorithms for power grid equipment failure, daily data, operating data intelligent analysis, by a large number of initial recording of data clean-up, according to the characteristics of safe operation of power factor extract and analysis linked to record data loaded into the data warehouse, and then process it accordingly mining algorithm to obtain the knowledge needed for the protection of the safe operation of the power grid to provide theoretical support.

II. DATA MINING TECHNOLOGY

Data mining is digging out from the mass of valuable data, technical information, the application range of technologies from a large number of, incomplete, noisy, fuzzy, random data extracted implicit in them, useful information and process knowledge [6]. Their work includes data integration, modeling, mining and knowledge analysis. Data can be structured, such as data in a relational database; it can be semi-structured, such as text, graphics, images, data, and even the distribution of electric power dispatch heterogeneous data between various services [7]. Data mining technology in electric power dispatching massive data processing and mining depth information has a great advantage, for the electricity dispatching business needs, you can use predictive model of knowledge discovery data mining, multidimensional analysis model, and correlation analysis models constructed mining model achieve knowledge extraction [8-9].

Run against massive amounts of data generated by smart grid, to tease out the common data classification technology, as shown in Figure 1, including: the decision tree algorithm, Bayesian classification and Bayesian network algorithms, neural networks, support vector machines, nearest neighbor classification algorithms [10].

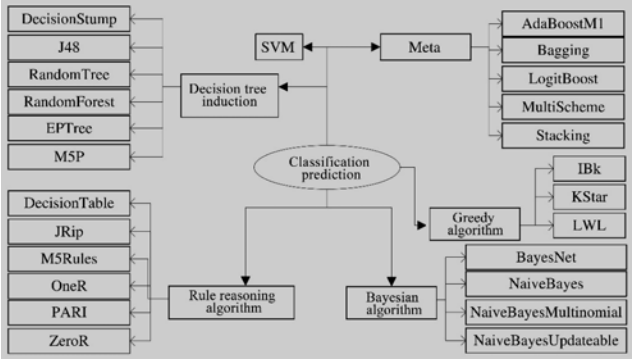

Figure 1.Architecture of classification method in smart grid.

Predictive models. In order to grasp the object of analysis of development, and the need to predict future trends using existing data sequence. Power load forecasting technique is commonly used in dispatch operations, in full consideration of system operating characteristics, capacity increase decision, the natural environment and social impact of the condition, research or the use of a system to be able to handle the load of mathematical methods in the past and the future, according to the temperature, humidity and other factors to predict future demand for electricity, to meet certain accuracy requirements of the premise, to determine the value of a particular moment of the load, guide power scheduling decisions. Load forecasting methods commonly used are: time series, gray prediction method, fuzzy clustering and prediction, neural network forecasting method and preferred combination forecasting method. Artificial neural network can establish any nonlinear models for time series forecasting problems to solve.

Multidimensional analysis model. Victoria is the level corresponding to the concept of information, multidimensional analysis to dimension, based on the statistical analysis of abstract data classification, according to analysis of the object's properties, characteristics, establishment of operational information classification model. Multi-dimensional analysis based on statistical theory, after the association between different dimensions for data analysis, usually by time (year, quarter, month, week, day) and regional (regional, provincial, city, district) for analysis Dispatching Analysis , commonly used methods of regression analysis, variance analysis. Electric power dispatching data classification model can also be used for statistical analysis, the data were summarized into three categories, extracting information representative of common characteristics.

Correlation Model. Association rules reflect dependence between things or related knowledge, association rule mining is focused on searching for relevance in a database or repository of project information collection or an object, related, or have information of causality, two important parameters involved are the minimum support and minimum confidence, support the probability rules pieces before and after the piece appeared centralized data simultaneously; trust said in pieces before the establishment of the rule after rule to launch probability member, after or before the rules with respect to the rules of member credibility. Correlation refers to a contact occurs when something else will happen, it can be described by the associated support and confidence. When power load characteristic analysis, the factors can change the load characteristics into two categories: one is the load factors have long-term impact effects, effects on long-term trends in loading performance of load changes, such as economic development, industrial structural changes; the other is the impact on the load factors have short-term effects, such as temperature, rainfall and other climatic factors. Research on the influence of various factors on the industry or regional electricity load change, help to improve the prediction accuracy of the power load, ensure the safety of the power system, economic operation.

III. SMART GRID DATA ONLINE STORAGE TECHNOLOGY BASED CLOUD COMPUTING

With the gradual advance smart grid construction, operating data and equipment online monitoring data in all aspects of the power system is recorded, mass data transfer and storage problems resulting not only cause great burden on the monitoring device, but also restricts the power system intelligent by leaps and bounds [11-13]. Through data compression can effectively reduce the amount of network data transmission and improve storage efficiency. Therefore, data compression technology gained widespread attention, discussed the compression and reconstruction algorithm based on lifting scheme of fault transient process signals in real-time data, the use of linear integer wavelet transform dual orthogonal filter combination Huffman coding method power system real-time data compression and decompression. It has been studied the lifting wavelet compression based on two-dimensional thermal power plants cyclical data algorithms. It has been studied the steady state power system data parameterization compression algorithm. Transmission line condition monitoring system in order to find insulator discharge, leakage current sampling frequency is relatively high, the amount of data. Currently such systems commonly used in wireless communication, network bandwidth is limited, hence the need for data compression. Data compression reduces the storage space on the one hand, on the other hand caused by compression and decompression consume a lot of CPU resources. After the data arrives monitoring center needs to decompress the data, we need a suitable computing and storage platforms.

In data storage, smart grid massive data distributed file system can be used to store, such as the use of Hadoop HDFS and other storage systems, but these systems although you can store large data, but it is difficult to meet real-time requirements of the power system. Therefore, the system must be based on the performance of large data and analysis requirements classification storage: very high performance requirements of real-time data using real-time database system; to the core business data using conventional parallel data warehouse system; a large number of historical and unstructured Data using a distributed file system. This paper presents smart grid build large multi-level data storage system, shown in Figure 2. It should be noted that, given the current cloud platform receives real-time smart grid

monitoring data is not guaranteed, you can set the number of front-end data access information integration front in FIG. 2 and is responsible for receiving the communication network in real-time alarm information sent or monitoring data, and is responsible for staging the cloud platform does not respond.

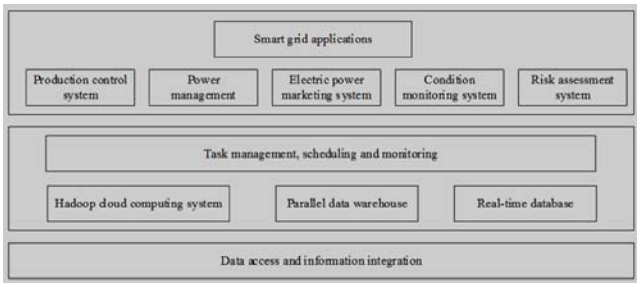

Figure 2. Multi-level storage system for smart grid.

In addition, the smart grid data formats and traditional business data has a very different, with its own characteristics. For example, in fault recording and transmission equipment condition monitoring, the waveform data more, and waveform data with traditional business data has a different nature, with data to generate speed and other characteristics. Therefore we need to study large data storage for smart grid format, thus facilitating subsequent data analysis and calculations. Smart grid environment under all kinds of heterogeneous data, cannot use the existing simple data structure to describe, and the computer algorithm is relatively inefficient in dealing with complex structured data, but the data processing is very efficient and homogeneous. Therefore, how to organize data into reasonably homogeneous structure, large data storage processing is an important issue. In addition, there are a lot of smart grid unstructured and semi-structured data, how these data into a structured format, is a major challenge.

In general, the larger the data, analysis and processing time will be longer. Traditional data storage solutions for the amount of data a certain size and design, within its design range processing speed can be very fast, but cannot meet the requirements of large data. Under the future smart grid environment, from power generation, transmission and distribution sectors, the electricity sectors, we need real-time data processing. Current cloud computing system can provide fast service, but there may be a brief network congestion, and even affect a single server failures, but not guaranteed response times. Memory-based database more attention. Memory Database is the data directly in the database memory operations. With respect to the disk, memory, data read and write speeds several orders of magnitude higher, the data stored in the memory than the visit can greatly improve the performance of applications from the disk. Current power system has begun to use in-memory database, in order to improve real-time. For example, some regions of our country with the electricity shortage last year, while the other parts of the state rendered surplus electricity, SAP launched HANA in-memory database based on smart meter analytical solutions [49], hoping to be involved in aspects of the smart grid and large power the user's data integration and consolidation analysis, in order to achieve the country's energy consumption analysis, in order to make the appropriate preventive measures. Focus on keywords in the large data query is also an important challenge. By scanning the entire data set to find the records that meet the requirements of the method it is obviously not feasible, even if a similar speed up the scanning Map Reduce such parallel processing techniques, as shown in FIG.

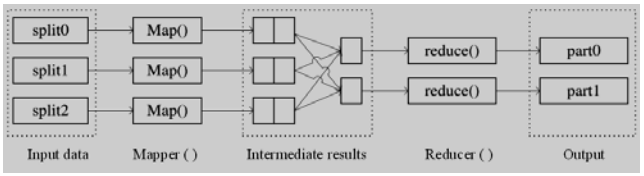

Figure 3. Parallel computing model of Map Reduce.

Through prior indexing structure data to help find is a more rapid method while saving system resources. The general index structure designed to support only some simple data types, big data is required to establish a suitable structure for the complex structure of the index data [50], which is a huge challenge. For example, the multi-dimensional data acquisition of things, the amount of data continues to grow, while the query time is required, the need to constantly update the index structure, the design of the index is very challenging. The following are the big data challenge the smart grid to bring in data processing from power generation, power transmission and electricity sectors analyzed.

IV. SMART GRID ONLINE DATA MINING SYSTEM BASED CLOUD COMPUTING

Cloud computing is an integrated system that will distributed computing and parallel computing and grid integration, based on the purpose of commercial implementations of constructed, you can achieve massive distributed data acquisition, storage, retrieval and parallel computing, has the characteristics of on-demand services, a variety of network devices and can, at different times and locations for a visit. Currently, Hadoop as the most typical, the most widely used cloud storage and cloud computing technology, the core part of the design comes from Google's distributed file systems, programming models and large-scale data management techniques. Data preprocessing refers to the core of the data analysis carried out before all the processing, smart grid, the data collection system to collect a variety of data to the original data. Data preprocessing can be divided into data cleaning, data integration, data conversion and data of the Statute of the four class. Smart Grid massive information processing, data preprocessing typically have a discrete Fourier transform (DFT), signal de-noising and data compression. Data mining covers multi-disciplinary knowledge, including databases, information retrieval, artificial intelligence, machine learning, and statistics and data visualization of the latest research results. Core data mining algorithms can be divided into three categories: classification and prediction, clustering technology and

association rules. As shown in Figure 4, shows the smart grid information processing massive data preprocessing and data mining basic relations and implementation ideas.

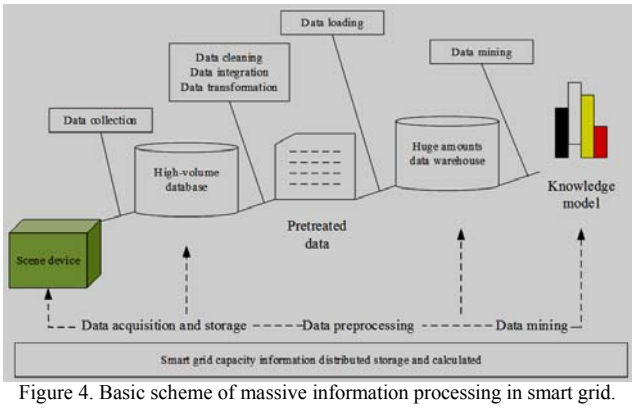

Figure 4. Basic scheme of massive information processing in smart grid.

The system's intelligent analysis mainly on historical data grid multidimensional correlation analysis and time series forecasting, reference value range of historical data generally valid for one year, longer and lose the reference value. Multidimensional data warehousing based relational store. Its architecture is schematically shown in Figure 5.

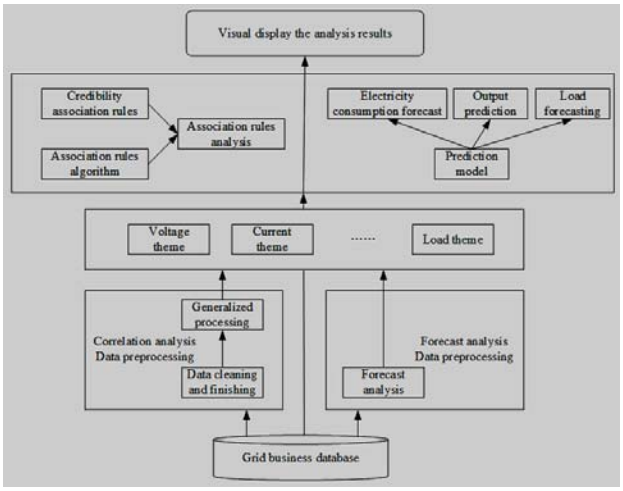

Figure 5. Architecture process of grid data smart analysis.

V. HADOOP SMART GRID FAULT INFORMATION PROCESSING SYSTEM BASED ON

Smart Grid fault information system faces several questions as follows: 1) a variety of devices generate large amounts of data in real time, the system requires a rapid response data; 2) data sources distributed across the entire network, you need to distributed storage and centralized management; 3) topology information unit is not in the whole network utilization; 4) data for the same event with a multi-source, thereby increasing the robustness of the need to integrate the decision-making mechanisms; 5) Data generated under different device and configuration parameters with heterogeneous characteristics, the need for a unified information modeling; how 6) Historical data migration and utilization.

Combined with Smart Grid fault information wide area, the main challenge panoramic, timeliness and mass and other characteristics, and fault information system Nowadays, proposed Hadoop-based smart grid fault information processing platform, the platform has data distributed acquisition module, cloud storage and cloud computing module, decision-making system and knowledge base module of three parts. Firstly, the basic framework of the platform, the architecture supports the following data mining and other advanced features; then discusses the problem Disturbance files are stored in Hadoop systems and efficient file retrieval Disturbance; further, by Parallel computing framework based on Map Reduce, fault events and fault information fusion extract; and finally, based on real-time digital simulation system physically mixed to typical four station system, for example, to verify the above-mentioned program, has proven its viability. In the smart grid, the data storage method from the analysis, the platform has good scalability and universality. As shown in Figure 6 Hadoop-based smart grid fault information system.

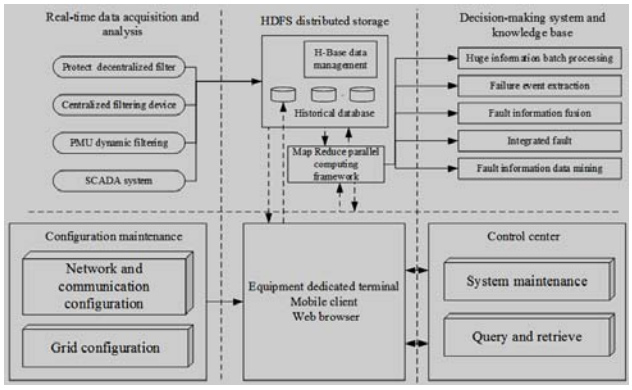

Figure 6. The smart grid fault information system based on Hadoop.

The system can be divided into distributed data acquisition modules, Hadoop cloud storage and cloud computing module, decision-making system and knowledge base module of three parts. Fault information system can record real-time acquisition wave from the protection, centralized record wave, PMU dynamic record wave data and SCADA systems, and scalable. HDFS sent to all the data on a distributed storage system, and to achieve effective management by H-Base data management techniques. Map Reduce parallel computing framework can be achieved by such information preprocessing, extraction of fault events, fault information fusion, fault synthetic discriminant functions. The fault information system has the following characteristics: (1) the grid all protection and fault recording devices IED management; the whole network collection (2) fault data, distributed storage and pretreatment, centralized Analytical; (3) a highly efficient data retrieval mechanism, feature extraction mechanism and heterogeneous information fusion function; (4) the automatic fault analysis, timely and accurate access to the fault location, type of fault

information; (5) the automatic analysis and evaluation relay protective equipment behavior.

VI. CONCLUSION

Future smart grids will be relying on large data processing and analysis techniques in real-time panorama grid. Cloud computing is such a heterogeneous and diverse data provides storage and analysis platform. Inevitably produce large data platform to run after a period, cloud and big data analytics platform will provide interworking state electrical equipment maintenance, self-healing grid, isolated information systems support, and become an important candidate, low cost, good system scalability (unlimited storage capacity), high reliability, parallel analysis and other advantages, there are a few cases in the international system is put into actual operation, but in real-time, data consistency, privacy and security, there is still a lot of challenges the need to find appropriate solutions. Processing large data is still lacking, yet people to explore. Business analysis system based on the scheduling of the techniques described herein can be constructed to meet the power scheduling function operating model and business needs, business analysis for dispatchers, dispatching operation to extract knowledge and information. Data mining technology is still faced with the challenge of the large amount of data, how to properly deal with redundant information and noise data mining results to enhance the effectiveness and practicality are the need to improve the information system construction elements.

REFERENCE

[1] Bonomi F, Milito R, Zhu J, et al. Fog computing and its role in the internet of things[C]//Proceedings of the first edition of the MCC workshop on Mobile cloud computing. ACM, 2012: 13-16.

[2] Wang D, Song Y, Zhu Y. Information platform of smart grid based on cloud computing[J]. Dianli Xitong Zidonghua(Automation of Electric Power Systems), 2010, 34(22): 7-12.

[3] Bughin J, Chui M, Manyika J. Clouds, big data, and smart assets: Ten tech-enabled business trends to watch[J]. McKinsey Quarterly, 2010, 56(1): 75-86.

[4] Simmhan Y, Kumbhare A G, Cao B, et al. An analysis of security and privacy issues in smart grid software architectures on clouds[C]//Cloud Computing (CLOUD), 2011 IEEE International Conference on. IEEE, 2011: 582-589.

[5] Han J, Kamber M, Pei J. Data mining: concepts and techniques: concepts and techniques[M]. Elsevier, 2011.

[6] Simmhan Y, Aman S, Cao B, et al. An informatics approach to demand response optimization in smart grids[J]. Natural Gas, 2011, 31: 60.

[7] Chen M, Mao S, Liu Y. Big data: A survey[J]. Mobile Networks and Applications, 2014, 19(2): 171-209.

[8] Simmhan Y, Aman S, Kumbhare A, et al. Cloud-based software platform for big data analytics in smart grids[J]. Computing in Science & Engineering, 2013, 15(4): 38-47.

[9] Simmhan Y, Giakkoupis M, Cao B, et al. On using cloud platforms in a software architecture for smart energy grids[C]//IEEE International Conference on Cloud Computing (CloudCom). 2010.

[10] Vermesan O, Friess P, Guillemin P, et al. Internet of things strategic research roadmap[J]. O. Vermesan, P. Friess, P. Guillemin, S. Gusmeroli, H. Sundmaeker, A. Bassi, et al., Internet of Things: Global Technological and Societal Trends, 2011, 1: 9-52.

[11] Yildiz M, Abawajy J, Ercan T, et al. A layered security approach for cloud computing infrastructure[C]//Pervasive Systems, Algorithms, and Networks (ISPAN), 2009 10th International Symposium on. IEEE, 2009: 763-767.

[12] Bin S, Yuan L, Xiaoyi W. Research on data mining models for the internet of things[C]//Image Analysis and Signal Processing (IASP), 2010 International Conference on. IEEE, 2010: 127-132.

[13] Fang X, Misra S, Xue G, et al. Smart grid—The new and improved power grid: A survey[J]. Communications Surveys & Tutorials, IEEE, 2012, 14(4): 944-980.