Big Data Management and Application Research in Power Load Forecasting and Power Transmission and Transformation Equipment Evaluation

Bo Hu\textsuperscript{1,3}, Chuan Pang\textsuperscript{1}, Liu Wang\textsuperscript{2}, Hanlu Chu\textsuperscript{2} and Chengjie Mao\textsuperscript{2}

\textsuperscript{1}Macau University of Science and Technology, Macau, China
\textsuperscript{2}South China Normal University, Guangzhou, China
\textsuperscript{3}Guangdong Power Grid Corporation, Guangzhou, China

Abstract. With the rapid development of the power grid, traditional power load forecasting and evaluation methods for power transmission and transformation equipment have become increasingly inapplicable. Accurate power load forecasting and evaluation of power transmission and transformation equipment are very important for grid planning and upgrading the quality of modern power systems. With the continuous emergence of new technologies, big data on smart grid is less systematic or practical. In this study, we propose framework for building a big data platform in the power grid industry, and elaborate the method to use big data to predict the level of power load and assess the status of power transmission and transformation equipment. Then, we deploy a practical exploration of big data application in the power grid industry. The proposed big data methods for forecasting and assessment on power grid revealed satisfactory results in practical application.

1. Introduction
We live and work in an unprecedented era of technological innovation. In the changing era of technology, enterprise must follow the trend and release the enormous potential which embodied in smart enterprise. With the rapid development of China's power industry, the scale of the power grid is expanding day by day. The accuracy of power load forecasting and the requirements for the safe operation of power transmission and transformation equipment are also increasing. The storage and analysis of massive power data flow will put forward higher and higher requirements for data processing of power information systems. Traditional data management and analysis will gradually no longer be applicable. With the help of big data technologies, the acquisition, processing, analysis and application of data information in the power industry will certainly activate the value contained in the power big data and tap the potential of the power big data market.

Power big data is the practice of big data concepts, technologies, and methods in the power industry. It involves the power generation, transmission, transformation, distribution, electricity, dispatch, and other links. It is a collection of data across units, disciplines and businesses. Power big data can be divided into three categories according to different sources: the first is electricity generation data from power generation enterprise; the second is the operation and management data of the state grid corporation of China, including data on transaction electricity prices, electricity sales, and ERP integration platforms; The third category is data on the power user side, including data collected by the user-side management platform and smart meters. Many domestic researchers perform power load forecasting and power transmission and transformation equipment evaluation under the framework of big data. However, many scholars pay more attention to explain the necessity and
importance of combining large data and grid application scenarios, and the advantages and disadvantages of different applications. There’s been a lot of talk about but rarely achieve [1].

The rest of the paper is organized as follows; Section 2 presents a framework based on big data environment, and gives how to use specific technologies to realize the research ideas of integrating big data with grid application scenarios. Section 3 elaborates the power load forecasting level and the state assessment of power transmission and transformation equipment. Section 4 provides operation level reference for big data technology application in power grid industry.

2. Digital Power Grid Big Data Platform
Digital power grid refers to a comprehensive information system represented by digitization, networking, intelligence and visualization on a physical structure and characteristic entity. It is not only for the internal service of the power industry, but also for all sectors of the society and the general public with strong characteristics of public welfare. It is the general name of power system, information system and innovative value-added service platform unified to realize digitalization. The core is informatization, automation and interaction.

Digital power grid is one of the important technology application areas of big data. The big data structure of digital power grid is complex and various with the dispensability, diversity and complexity. These characteristics pose great challenges to big data processing. The digital power grid big data platform is the basis of big data mining. Through the digital power grid big data platform, the entire data sharing of the power grid can be realized, and it also can provide support for the development and operation of business applications.

The big data application of digital power grid needs to be built on the big data platform, which provides unified data access, cleaning, storage, management, analysis and calculation for the application. The core platform of big data consists of basic resources, data storage, data analysis and processing, data management, platform service and platform control. The big data platform architecture of digital power grid is shown in figure 1.
Figure 1. Big data platform architecture of digital power grid.

Combined with the application requirements of digital power grid, various kinds of big data applications can be constructed on the core platform of big data. Big data management and application research is carried out in the following two aspects: electric load forecasting and power transmission and transformation equipment status assessment.

3. Electric Load Forecasting

The theory of power system analysis shows that maintaining the balance of supply and demand is related to the safety and stability of power usage in the power system. Under the premise of ensuring the safe and stable operation of the power system, the economy of operation should be improved. Accurate power load forecasting is the premise of power system dispatching, real-time control, operation planning and development planning. It is the basic information that the power grid dispatching department and planning department, and it is of great significance to formulate reasonable planning and scheduling schemes, improve equipment utilization, and guarantee safe and stable operation of the system.
3.1. Traditional Power Load Prediction Method

User load forecasting research based on power data is divided into short-term load forecasting and mid-long term load forecasting [2]. The short-term power load forecast refers to the forecast of power load in the next few hours or a few days. And it has been used as an important functional module of the energy management system to provide a basis for arranging power dispatch plans, power supply plans, etc. Medium-term load forecast refers to the load forecast in the next year, and the forecast result serves as an important basis for the unit maintenance plan and power grid planning. Long-term load forecasting refers to the forecast of electricity load in the next 1 to 10 years, which provides reference for the planning, expansion and expansion of power grids.

At present, the theory of short-term load forecasting has gradually matured[3], including regression analysis method, time series method, wavelet analysis, support vector machine, artificial neural network, fuzzy prediction, comprehensive model prediction and other methods. Compared with short-term forecasting, medium and long-term load forecasting is more susceptible to uncertainties, such as weather conditions, natural environments, and human activities. Therefore, medium and long terms load forecasting require more data and also are more difficult. Mid-long term load forecasting methods can be divided into two categories: parametric model-based methods and non-parametric model-based methods. The medium and long term forecasting methods based on the parametric model include the electricity elasticity coefficient method, the time series method, and the correlation analysis method [4]. The methods based on non-parametric models include grey prediction techniques. No matter what kind of forecasting method, the core problem is to build a forecasting model based on historical load data of power users. The accuracy of the model determines the level of the forecasting.

In recent years, with the rapid deployment and development of smart grids in China, it has brought direct challenges to power load forecasting. The data of various types of sensors and smart devices has continuously increased and the data obtained from the devices and the relevant factors of power load predictions collected by various sensors, such as temperature, weather, and wind speed, has increased dramatically. The data dimension also continues to increase, and the data size also grows from GB to TB or even higher. Traditional load forecasting methods can no longer meet the requirements of efficiency and accuracy in load forecasting, the domestic and foreign scholars have focused on the load forecasting based on the technology of large data forecast.

3.2. Power Load Forecasting Based on Big Data

The power load is affected by factors such as meteorological conditions, economy and user's electricity behaviour. The factors make it difficult to achieve high-precision load forecasting. The power load forecasting method adopted in this paper is to use the historical data and current data of power load and data mining algorithms to mine the sensitive factors affecting load fluctuations [5]. This paper also establishes a prediction model under big data environment which can forecast all kinds of power loads. The framework is shown in figure 2.
Power load forecasting systems based on big data architecture can be divided into three levels:

(1) Application layer. The load forecasting results are applied to power grid scheduling, distribution network planning and other fields.

(2) Prediction analysis layer. Analysing the sensitive components and associated relationships that affect the load based on various types of data, and various types of load forecasting are performed through the prediction model and calculation processing based on big data.

(3) Data integration layer. Integrate internal and external data, such as grid dispatching load, user electricity consumption and economic society data of providing basic data for load forecasting.

Specifically, Spark clusters can be used to build big data basic storage platforms [6], which integrate the data collected by the grid subsystems into big data storage, and use the parallel computing framework to rapidly mine and analyze smart grid big data, migrate traditional load forecasting to cloud computing platform and use the random forest algorithm to achieve parallel load forecasting. That is based on the Spark platform and parallel random forest regression algorithm for power load forecasting. The parallel random forest regression load forecasting process based on the Spark platform is: raw data loading--data cleaning and pre-processing--a parallel random forest regression model establishment--spark large data processing platform--comparison and evaluation of load forecast results--end. The Spark distributed computing platform and the parallel random forest regression algorithm are introduced as follows.

3.2.1. Memory-based distributed computing platform - Spark. Spark is a distributed clustering framework based on memory operations, designed to simplify the writing of parallel programs on the cluster. In the cluster, a high-memory task scheduling server acts as a master driver, and a number of computers with a large hard disk space serve as worker. In the calculation, the driver accepts user instructions and tasks, and divides the task for each worker to execute. The worker can extract data from the Hadoop distributed file system or other distributed file system when executing the task, and stores the calculated data in memory, then returns it to the driver. The driver combines the data returned by each worker to get the final result. Spark inherits the linear scalability and fault-tolerance of MapReduce and improves the strict implementation process of MapReduce that must map first and then reduce. Spark can directly transfer the intermediate result to the job through a directed acyclic graph operator, rather than storing them in a distributed file system like MapReduce.

3.2.2. Parallel random forest regression algorithm in distributed environment. Random forest algorithm is a common combinatorial algorithm, the principle is to build a set of base classifiers from

---

**Figure 2.** Power load forecasting framework based on big data architecture.
training data and vote on the prediction results of each base classifier to decide the classification. The effect of the combined classifier is often better than that of the base classifier. The random forest algorithm is composed of multiple decision trees that form the “forest” and its base classifier is the decision tree. Comparing to the single decision tree algorithm, the random forest algorithm has better effects of classification and prediction and less overfit, because the construction of the random forest algorithm has two characteristics:

(1) The training set composition of decision trees in random forests. The method of forming a sample set of the same size (i=1,2,...,k) by k random extractions of equal samples was used. Because of randomly selected, for each of the k decision trees, the training set of each decision tree is different and statistically equivalent of the random sampling form the data set, thus it retained the characteristics as much as possible.

(2) The growth of a single decision tree. The key to the growth of the decision tree is the choice of the "optimal splitting attribute." In this algorithm, m(m≤M, M is the total number of attributes of the sample) attributes are randomly selected for each splitting node as the splitting attributes to be judged, and randomness is introduced into the growing process of the tree.

Due to the two above characteristics, the random forest algorithm makes the model well retained the original statistical characteristic of data set and prevent the possible over-fitting in the process of model training.

Therefore, on the basis of summarizing the traditional single-machine power load forecasting model, we can combine the research on the Spark distributed computing framework to implement the parallel random forest regression algorithm under the Spark platform, furthermore, a short-term power load forecasting method based on Spark platform and parallel random forest regression algorithm is proposed, which can well deal with the requirements of high-precision load forecasting and mass data computing in the context of current power big data. According to laboratory test data, the accuracy of short-term load forecasting is 20%~30% higher than that of traditional forecasting methods and the long-term load is increased by 15%~20%. It is of great significance to adjust the load of heavy-load lines to increase the transformer capacity and eliminate the load supply gaps in the area of unbalanced power load.

4. Status Assessment of Power Transmission and Transformation Equipment

With the continuous development of modern power technology, there are more and more monitoring methods for the operation status of the main equipment, including live detection, on-line monitoring, bad working conditions, and historical events. This amount of information reflects the internal and external operating status of the main equipment of the power grid. It can be used to analyze the operational status of the equipment and make decision on equipment maintenance. The big data is an extremely important tool which can be used for data mining and analysis to support the actual grid production operations [7].

4.1. Traditional Status Assessment Method for Power Transmission Equipment

At present, China's power grid companies use a wide range of power transmission and transformation equipment status assessment methods, including equipment status scoring system, expert system method, multi-dimensional equipment status evaluation method based on traditional machine learning, and sample training methods that introduce remote expert opinions [8].

However, in recent years, with the development of intelligent monitoring equipment, the amount of state parameter data of power transmission and transformation equipment is increasing exponentially, and the equipment state data comes from many different systems. Traditional state assessment methods cannot handle such multi-source heterogeneous mass data. On the one hand, the elements of the original status evaluation system are limited to the status information of some equipment such as defects, experimental data, and bad working conditions, and fail to incorporate all the information amount reflecting the equipment status into the evaluation system; on the other hand, the original state assessment method use the decision tree and the expert score, set the weight and the deduction component of each situation, so as to comprehensively determine whether the state of the equipment is
normal, needs attention or exceptions, etc [9]. Thus, the research on state assessment and risk assessment of power transmission and transformation equipment based on big data technology is of great significance for the comprehensive evaluation and risk assessment of the internal and external operating conditions of the main power grid equipment.

Therefore, we can develop key technologies for power transmission and transformation equipment based on big data, such as load capacity dynamic assessment, fault prediction, state assessment and operational risk assessment of power transmission equipment, next to develop cross-platform data acquisition and conversion device, and then a comprehensive analysis system integrates information from the power grid, equipment and environment information. As a result, it can control the state of transmission and transformation equipment accurately and comprehensively in real time, and provide technical support for the state maintenance of transmission and transformation equipment.

4.2. Status Assessment of Power Transmission and Transformation Equipment Based on Big Data

The establishment of quantitative evaluation indicators for state assessment and risk assessment of transmission and distribution equipment based on big data technology belong to the category of decision support systems, which is based on the external system to provide raw data [10]. By continuously adapting to the system's rapidly changing evaluation indicators and advanced assessment methods, the state assessment and risk assessment of power transmission and transformation equipment can be achieved. At the same time, the system should also be combined with various intelligent detection modules to form unified communication channel to achieve data sharing [11]. The overall system framework for state assessment and risk assessment of power transmission and transformation equipment based on big data technology is shown in figure 3.

![Figure 3. Overall framework of state assessment and risk assessment system for power transmission and transformation equipment based on big data technology.](image)

The key technology of equipment state evaluation based on big data analysis can be applied from the aspects of distributed storage, frequent item mining technology and clustering algorithm technology.

4.2.1 Application of distributed storage in equipment state evaluation. It is generally believed that big data can deal with much more data than traditional relational databases. In terms of storage, some NoSQL databases broke through relational databases emerged, such as MongoDB, Hbase, etc. Because of distributed databases store data on different machines, they require additional communication and management overhead compared to relational databases [12]. Distributed
databases need to solve the problem of data consistency and performance. In distributed database, the problems of consistency, availability and fault tolerance cannot be satisfied simultaneously. When faced with different problems, it is necessary to select a processing strategy according to the focus of the problem. In general, it contains three points: (a) to avoid single point of failure resulting in data loss and more data needs to be backed up; (b) if there are more backup data, more data synchronization is needed to ensure the consistency of the data; (c) if the data consistency requirement is high, the performance is reduced, which increase the amount of additional communication and management overhead. At present, in the power system, the status information generated by various devices through daily monitoring has exceeded the PB level. These data are very large, and the types of devices vary from device to device. Therefore, using a distributed database to store device status data is very necessary.

**4.2.2 Application of frequent items mining in equipment state evaluation.** Association rules excavates the relationship of valuable data items from a large amount of data. In practice, it is often used to find static laws in disorganized data. The common association rule algorithms include Apriori and FP-growth, etc [13]. The idea of the Fp-growth algorithm is as follows: first, scan all the data, generate all the candidate frequent item sets, and sort them in descending order to cut off the elements whose support is lower than the threshold; Secondly, scan all the data again and generate the FP tree according to the result set of the first step. Finally, from the FP tree, find meaningful frequent items according to the rules. In the equipment status assessment, association rules can be used to discover the factors related to the equipment status. Divide the status of the equipment into different levels, and record an equipment level and its corresponding discrete weather factors, geographic factors, traffic factors, human factors, etc. Establishing FP-growth tree and discovering the tree whether the "serious state” has frequent items. If frequent items are found, finding out the corresponding factors of “serious status” from frequent items and excluding them to maintain the safe operation of the power grid.

**4.2.3 Application of clustering algorithm technology in equipment condition evaluation.** The method of equipment fault diagnosis based on big data mining is to mine the common fault modes of the equipment through the cluster analysis of fault abnormal state parameters. One of the most mature methods is the k-means clustering method. The algorithm gathers the object near its closest center point. If you want to determine the center point, you need to know which objects are included. In the k-means clustering, the selection of k-values of clustering clusters is very important. Only by finding the appropriate k-values can we get a more ideal clustering effect. The contour coefficient can solve this problem well. It combines the concepts of agglomeration degree and separation degree and is more effective to evaluate the clustering effect through the contour coefficient. The measure of the degree of aggregation in a cluster is to calculate the average distance between the i-th element in the cluster and other elements in the cluster, which is denoted as \( a_i \). For the quantification of the degree of separation among clusters, the method is to select a cluster b other than the cluster of the above elements, calculate the average value of the distance between the element and all the other elements in b, and then calculate the distance between the above element and all elements in the cluster where the element is not located, and find the minimum value of the distance between the element and other clusters, denoted as \( b_i \). The calculation formula for the i-th element contour coefficient is:

\[
S_i = \frac{(b_i - a_i)}{\max(a_i, b_i)}
\]

Finally, the contour coefficients of all elements in all clusters are calculated, and the average of the contour coefficients of each element is calculated as the overall contour factor of the current cluster.

Frequent item mining, and clustering algorithms research, the status assessment of power transmission and transformation equipment under the big data platform can be further realized, based on the technologies of distributed storage, which can ensure the actual assessment work implemented more safely and accurately, and the reliable operation of the power transmission and transformation equipment and the power system. According to the experimental test data, 21% of the blackout time can be reduced when compared to the traditional method, and in terms of increasing the availability of
assets, the company has increased from the original three nines to four nines, and the maintenance costs can be reduced by 28%.

5. Summary
With the development of society and advancement of technology, the current domestic level of economic development is gradually increasing. In order to better meet people's needs for the power grid industry, we must combine the enlightenment that big data brings to enterprises and constantly promote the development of the power grid industry to meet people's needs for the actual operation of the power grid industry, and promote digital transformation of enterprises and maximize the development of social economy. The research in this paper can provide reference for the application of big data technology in power load forecasting and state assessment of power transmission and transformation equipment.

6. Acknowledgments
This work was supported by the Major Science and Technology projects of Guangzhou, China (201704020203), and the National Science and Technology Support Program (2012BAH27F05).

7. References
[1] Tangjie Liang. Research on Power Load Forecasting Method and Its Application [J]. Communication World, 2018 (03): 238-240
[2] Ye Zhang. Application of Big Data Technology in Power Load Forecasting [J]. China Hi-tech, 2017(1)(07): 76-78
[3] Jiamei Wu. A review of research on short-term power load forecasting methods [J]. Modern Commerce, 2018 (08):176-177
[4] Guojian Chen, Qing Chen, Xiaotian Su. Study on medium-term power load forecasting and load model [J]. Modern Manufacturing Technology and Equipment, 2017(11): 59-60
[5] Lifang Chen. Research on Application of Power Forecasting System for Big Data [J]. Guangdong Science and Technology, 2017,26(02):89-90
[6] Dandan Li. Application of Spark Technology in Power System [A]. China Electric Engineering Society Electric Power Informatization Committee, 2017 Power Industry Informatization Conference Proceedings [C]. China Electric Engineering Society Professional Committee for Electric Power Information, 2017: 4
[7] Xiaonong Yu. Application of Big Data Mining Technology in Fault Diagnosis of Power Transmission and Transformation Equipment [J]. Communication World, 2017 (22): 151-152
[8] Jingde Chen, Gehao Shen, Jijian Wu, Yougang Xu, Fuji Wang. Application Status and Prospect of Big Data Technology in
[9] Zhong Zhen. Application of Power Load Forecasting in Smart Grid [J]. Quality Management of Fujian, 2016 (04): 175.
[10] Yanwei Dong, Yuyan Man, Lei Wang, Chuanghua Liu, Congli Liu. Research on Condition Evaluation and Risk Assessment of Power Transmission Equipment Based on Big Data Technology [J]. Electrical Application, 2017, 36 (22):78-83
[11] Guang Zhou, Danfeng Yan, Keguang Xu, Sun Li. Research on Big Data in State Assessment of Power Transmission Equipment [J]. Software, 2016, 37(01): 9-13
[12] Chen H. Research on Post Evaluation Index System and Marking Method for Electric Transmission and Transformation Equipment Maintenance Project [J]. Guangdong Electric Power, 2012
[13] Jun Hu, Liguan Yin, Zhen Li, Lijuan Guo, Lian Duan, Yubo Zhang. Fault Diagnosis Method of Power Transmission Equipment Based on Big Data Mining Technology [J]. High Voltage Technology, 2017,43 (11): 3690-3697
[14] Enmin Li. Research and Implementation of State Maintenance Technology System for Power Transmission Equipment [J]. Technological Innovation and Application, 2016 (14): 65-66
[15] Xun Wang, Xin Wang, Meng Zhao, Peng Xu, Lin Ma. Discussion on Application of Big Data
Analysis in Power Transmission Equipment State[J]. Power Big Data, 2018, 21(01): 1-5

[16] Lifeng Lin. Application of Power Load Forecasting in Smart Grid [J]. Electromechanical Information, 2016(03): 17-17.

[17] Bowen Zhang, Chunyu Yan, Jiangang Bi, Feng Wang, Shuai Han. Research on the Structure of Transmission and Transformer Equipment Condition Warning System Based on Big Data [J]. Power Information and Communication Technology, 2016, 14(12): 26-32

[18] Xi Luo, Wentao Shen, Mingwei Xie, Shuhang Yao, Shenjia Tang. Key Technologies and Development Trends of Intelligentized Transmission and Distribution Equipment [J]. China High-tech Zone, 2017 (22): 141

[19] Li M, Lu W, Xiang D. Study on Power Transmission and Transformation Equipment Intelligent Operation and Maintenance System and Its Application [J]. Journal of Electrical Engineering, 2015