Distributed Data Mine OSGI Technology for Electrical Supply Smart Grid

A. Srinish Reddy[1] J. Manikandan[2] Goddeti Mallikarjun[3]
[1,2,3]Department of Computer Science & Engineering
[1,2,3]St. Martins Engineering College, India

Abstract— With the increase in technology and data every day utilization, there has been a tremendous growth in the amount of data generated. Traditional analysis system has bottlenecks of performance and scalability in big data processing. The research and development of novel and efficient big data analysis and mining platform has become the focus of all organizations. Along with the development of smart grid, power data with characteristics of power industry needs more targeted and efficient data mining analysis. In this paper, aiming at the shortage of existing work, we propose a distributed big data mining platform based on distributed system infrastructure such as Hadoop and Spark. The platform develops and implements a variety of rapid highly parallel mining algorithm by Spark and Tensorflow, including machine learning, statistics and analysis, deep learning and so on. Using the OSGI technology to build low coupling component model, the platform improve reusability of component algorithm, introduces the workflow engine and user-friendly GUI, reduces the complexity of the user operations, support user-defined data mining tasks. For the characteristics of smart grid big data, the platform develops and improves the dozens of algorithm components about data processing and analysis. And designing a scalable algorithms library and the component library greatly improves the scalability of big data mining platform and processing smart grid data. Our platform has already been launched in a state grid Company, satisfying the demand of various smart grid data analysis business.

Keywords: Parallel; Data Mining; Components; Spark; Workflow

I. INTRODUCTION

Along with the rapid development of modern information technology and explosive growth of global data, large data has become the most important impetus for nations, enterprises and even the efficiently sustainable development of society. The world has entered the era of big data. Effectively analyzing data can't lack the support of data processing tools and machine learning platforms. Traditional analytical system is based on OLAP(Online Analytical Processing) system and OLTP(Online Transaction Processing) system. These systems perform quite well on the process of data analysis. Due to the limitation of the stand-alone operation mode, the data processing at the big data level reveals defects such as long processing time and insufficient performance. Since the inception of distributed systems infrastructure 1.0.0 formal version of Hadoop in 2011, distributed processing systems, represented by Hadoop and Spark, has drawn the attention of researchers and been widely used. The data mining platforms have achieved the transition from the data analysis to big data mining. With the excellent algorithm performance exhibited by deep learning in recent years, as well as the outstanding performance in image analysis, speech recognition and target detection. The new goal of data mining analysis systems is to apply and integrate deep learning algorithms.

At present, based on distributed computing frameworks such as Hadoop and Spark, and deep learning frameworks such as Tensorflow and Caffee, various general-purpose big data analysis and mining platforms have been developed. While the algorithms implemented by the analysis algorithm library are aimed at general-purpose data and lack of the abilities to make good use of characteristics for industry data. In the face of industry professional data, it is difficult to deal with frequent changes in complex application business and decision-making needs. Therefore, the development of big data mining platform for domain business data has become the focus of further research.

Power big data resources are similar but different from other business data. They are large and complex data resources, including internal data such as power grid operation data, equipment inspection data, enterprise marketing data, power enterprise management data, and external data related to grid data, such as weather data, national economic operation data, etc. To maximize the potential value of power data resources and make full use of these massive power data resources, it is necessary to obtain comprehensive information for power resource allocation decisions through comprehensive data analysis and industry characteristics analysis. Moreover, power data analysis involves many aspects of power production operations. People in different departments often pay attention to specific analysis functions, resulting in the need to generate a large number of analysis modules. Chinese existing power data analysis platform has better performance for targeted processing of power business, but the speed of data mining analysis and processing due to technical framework cannot meet the current situation of rapidly increasing power big data.

This article aims to solve the deficiency of the major data analysis platform. Based on a Spark, Hadoop, YARN and other frameworks, it proposes the distributed data mining platform that oriented to the power of big data. The platform integrates deep learning distributed algorithms and also supports data statistical analysis and traditional machine learning. The platform introduces workflow mechanism to provide a DAG (Directed Acyclic Graph) abstract data flow for big data analysis process, which reduces the development and the complexity of the user operation and increases the degree of component reuse. Modular building by OSGI technology are adopted to decrease the coupling between modules and simplify the operations of function expansion. At the same time, in view of the complex application of the electric power industry business professional, the platform integrates and improves the corresponding data analysis algorithm for power data. Besides, for supporting frequently

changing decision needs, it provides the easy expanded modular structure and algorithms library expanded interface.

II. PARALLEL DEEP LEARNING NETWORK

A. Deep Learning Network

The basic unit in a neural network is a neuron model, which includes input, output and computational functions. And a neuron model is expressed as the (1).

$$Y_j = f\left(\sum_{i=1}^{n}(X_i + W_{ij}) + B_j\right) - - - -(1)$$

Where Y_j represents the jth output result of the neuron model, X_i represents ith input element of the neuron model, W_{ij} represents the product of X_i and the weight of the jth neuron and B_j represents the bias of the jth neuron. A plurality of neurons is combined to form a level in a neural network structure, and a plurality of layers is stacked to form a specific neural network. With pre-training method proposed by Hinton to alleviate the local optimal solution problem in neural networks [12], the hidden layer is deepened to 7 layers, and the neural network is promoted to a true deep neural network. Deep neural networks often use a fully connected form to connect the lower neurons and the upper neurons, which will lead to excessive expansion of the number of parameters, and it is easy to fall into the local optimum. The CNN (Convolutional Neural Network) [13], which can reduce the number of free parameters in the network, becomes a more suitable neural network structure. CNN is inspired by the study of visual cortical electrophysiology in biology. By introducing a convolutional layer, CNN uses the convolution kernel as an intermediary between neurons and shares parameter weights, which greatly simplifies the model complexity and reduces the parameters of the model.

At the same time, the deep neural network can't model the changes at the time series level, and the accuracy of the application of natural language processing, speech recognition and other time series data is relatively low. The RNN (Recurrent Neural Network) acquires historical time characteristic information by applying the output of the neuron as another input signal to the next time stamp.

B. LeNet-5 and LSTM Parallelization

The LeNet-5 network proposed by Y.LeCun [2] is one of the most classic convolutional neural networks with 7 layers. Each layer contains trainable connection weights and adopts a strategy of weight sharing between each layer. It requires multiple rounds of iterative calculation, and the sharing of parameter weights between each layer provides basic conditions and optimization space for distributed computing. And the long-distance neuronal spacing in RNN will result in the long-term dependence problem, which cannot learn and use remote information. Based on the unit improvement in the original RNN, the LSTM (Long Short-Term Memory) network structure [8] remembers long-term information, and can avoid the long-term dependency problem in RNN.

This paper extracts the distributed computing engine Spark based on memory computing and Tensorflow as the implementation framework of LeNet-5 and LSTM networks. It can accelerate the training with the help of distributed computing features, realize the training parameters to improve the training precision and reduce the training loss. In each round of training, one node in the Spark platform is selected as the training parameter server, and other computing nodes perform model training by obtaining the obtained data fragments to obtain the model parameter variation __. The parameter server will receive the parameter variation calculated by each computing node, update the model parameters and the copy of the model parameters in each computing node, and perform a new round of training until the final training is completed.

The deep neural network model parallelization implementation algorithm is shown in the Algorithm 1:

1) Algorithm 1 Parallel Deep Neural Network Model Training Based on Spark

Input: Training data set

Output: Lenet-5/LSTM network model for training

1) Obtain the training data set, determine the data fragment size and the number of partitions, and initialize the model training parameter _:
2) Distribute data fragments and model training parameter copies to each compute node.
3) Each computing node extracts one of the allocated data fragments for network model training
4) The parameter server receives the model parameter variation of each computing node.
5) Update the central model parameters and the copy of the model parameters in each compute node, and adjust the network model using the back propagation mechanism.
6) Repeat step 2 until the training is complete.

The Fig.1 shows a schematic diagram of training based on Spark parallelization LeNet-5 network model

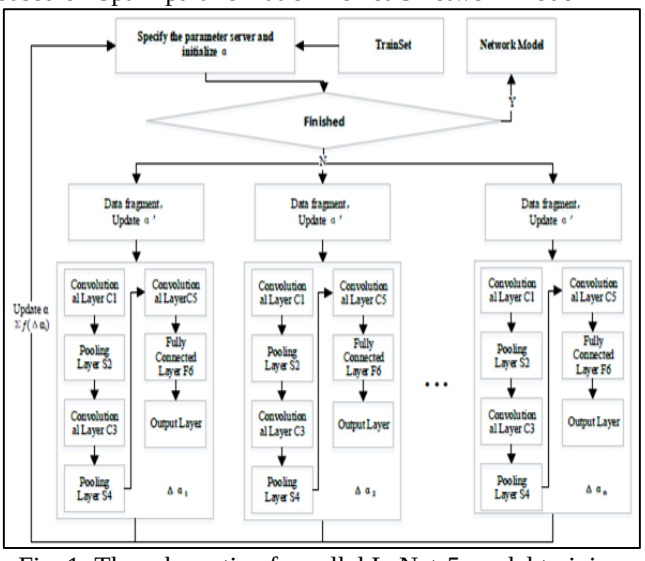

Fig. 1: The schematic of parallel LeNet-5 model training based on Spark.

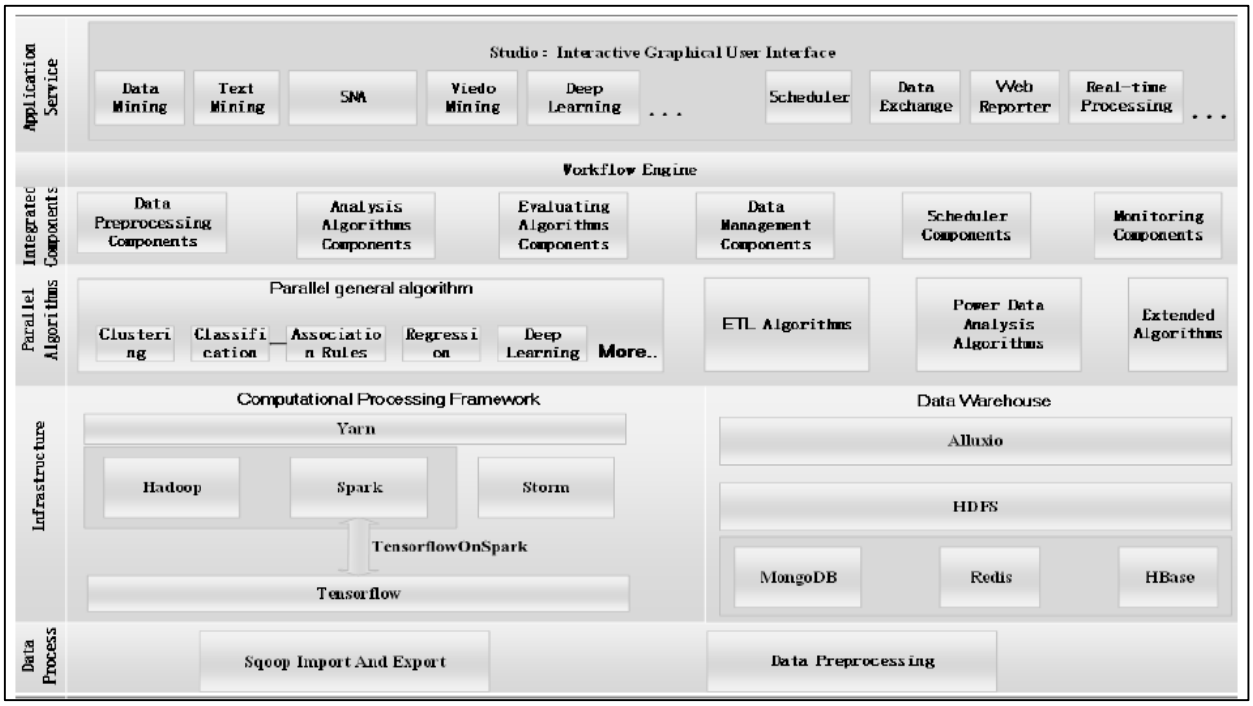

Fig. 2: The overview and architecture of the platform.

III. PLATFORM OVERVIEW

The distributed parallel data mining platform for power big data proposed in this paper adopts a scalable and easy-to expand five-layer architecture, as shown in Fig.2. From the aspects of data visualization, data management, and execution monitoring, it deconstructs the characteristics of power big data analysis and provides effective and convenient analysis services.

A. Data Process

The data process layer is aimed at raw power data originating from different data sources, including such as GIS data and EMS data. To process these various large data, the parallel platform adopts the MapReduce and implements more than 50 parallel distributed ETL algorithms, which is robust and more efficient to meet various types of data processing demands.

At the same time, tools, such as Sqoop, are provided to transfer the data extracted in the original data system to the distributed storage System. And it also improves storage performance and increase storage security by using the strategy of multiple backups.

B. Infrastructure

The infrastructure of platform consists of two module parts, one of which is the data warehouse module consisting of NoSQL, HDFS, Alluxio, etc. The data warehouse module stores structured and unstructured data in data partitions on disks. For the data stored in the data warehouse module, the parallel platform defines unified metadata information. The metadata information is composed of data storage type, data storage location, data storage amount and other information. To support the whole data warehouse module, the metadata is also stored in HDFS.

To meet demand, quickly reading and writing the data, of the various algorithms in the upper structure, the data warehouse module also combines the memory distributed file system Alluxio with HDFS to provide data storage in memory or other storage facilities in the form of files. The service provides a reliable data sharing layer for the upper distributed computing framework, while reducing redundant storage and resource recovery time.

The other module is the computational processing framework module, which implements a multi-hybrid computing framework and is composed by Spark, Hadoop, Tensorflow, etc.

Each computing framework has its own different resource management systems. In order to realize the overall management scheduling of the hybrid computing framework, the parallel platform realizes the unified resource management module based on the YARN resource management framework of Hadoop. At the same time, the TensorflowOnSpark framework is introduced to assist the docking of the Tensorflow framework and the Spark computing framework. The module also provides unified resources for the upper layer application and avoids conflicts between resource allocations.

C. Parallel Algorithms

The parallel algorithm layer is the core of the distributed data parallel mining platform. By improving the parallelism of the calculation in these algorithms and designing new calculation process in algorithms, dozens of parallels algorithms with high degree of computation and high computational efficiency are developed. And it also implements the commonly used distributed parallel machine learning algorithms and various types of ETL algorithm for data preprocessing modules. At the same time, commonly used deep learning algorithms, such as CNN, RNN, LSTM network and Bi-LSTM network, are also implemented. The coupling degree between the algorithms is low, and the algorithm uses the reserved interfaces of other algorithms to

call up. It is convenient for the algorithm to improve the operation by the method.

For the special services involved in the power system data and the unique features of power data, the general data mining analysis algorithms are difficult to do the trick. To solve the multi-timing, complexity and partial professional power equipment characteristics of the power data, more than 20 special algorithms have been developed and added into the platform, such as chromatographic differential warning and time series prediction. At the same time, an algorithm expansion interface is provided. The algorithm library can append the targeted algorithm and develop the original algorithm for the requirements of the business service.

D. Integrated Components

Industry business applications and decision-making needs often change frequently. And the dimensions of different business mining analysis about the data are not the same. The dimension of the power data mining analysis is accompanied by complicated power service characteristics. For example, transformer oil chromatographic data in power data, which contains relevant gas content data such as CH4 and O2. What the non-power industry personnel maybe pay attention to is the curve and trend of each gas content, while the power industry personnel pay attention to the analysis and excavation whether the amount of some important gases, such as CH4, CO2, H2, exceeds the threshold and each of them correlation between gas content and transformer failure.

For the complex demands, the parallel platform builds up a component-based, service-oriented development mechanism and runtime environment in the Fig.3. It decomposes the required development functions into multiple component sets. The information of component sets will be transferred in the form of DAG. The workflow engine in the platform will analyze the DAG and uses the OSGI (Open Service Gateway Initiative) service to execute scheduling operations, such as start-stop, update, and uninstallation, of the corresponding functional components. Depend on the mechanism; the business function application is highly dynamic.

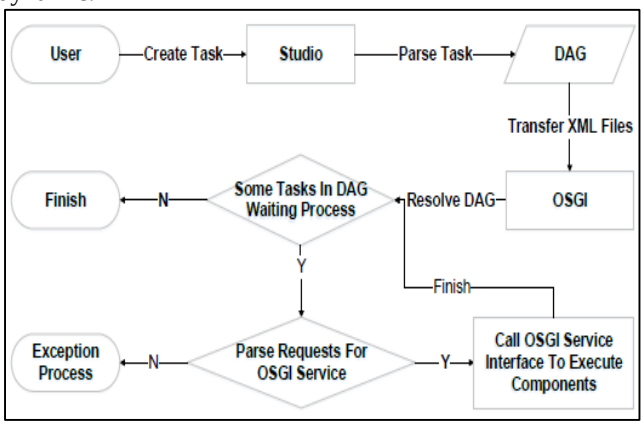

Fig. 3: The flow chart of scheduling components.

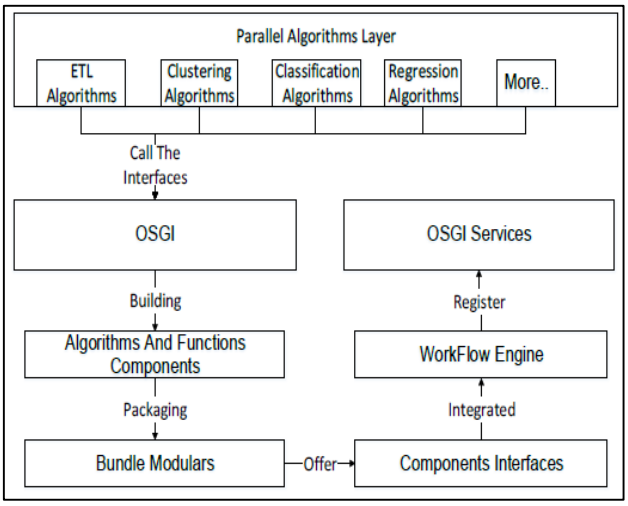

Fig. 4: The flow chart of integrating components.

Each component calls the interface of the parallel algorithm layer and obtains the relevant algorithm to implement the operating logic of component. It adopts the OSGI framework to enrich the corresponding operating logic and encapsulate it into the most basic module form, which is named bundle. It also provides interface for calling up and is integrated into the workflow engine. The flow is shown in the Fig.4. Each integrated component provides some interfaces, which could update and expand the functions of components, supported by OSGI technology for administer. The administer could modify the original components according to the business requirements without affecting other parts, and improve some operating logic to support the business analysis.

IV. THE FAULT TEXT OF POWER TRANSMISSION AND TRANSFORMATION

Analysis the parallel distributed power big data mining platform is a cloud computing web service application. The user can access and operate the parallel platform through the browser to execute the data analysis, and the computing process is at the cloud computing service node.

The user interacts through the interactive interface Studio of the parallel platform. The Studio is as shown in the Fig.5. The operations of users are constructed in the form of a workflow. The workflow is constructed in the form of the data flow diagram that the component is a node and the data interaction between the components is connected to the node.

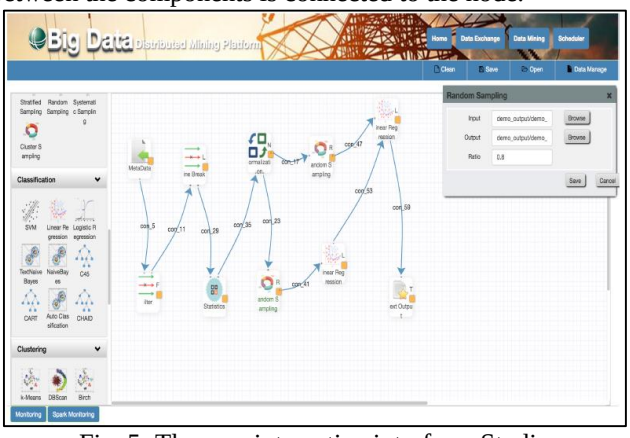

Fig. 5: The user interactive interface: Studio

The workflow treats the entire data mining analysis process as data flowing and converting in a data channel. Each time the data flows through a component node, it is converted into corresponding data. If a node in the data channel fails or the data fails to be processed, the data flow no longer flows to the subsequent flow, and the cause of the task failure is prompted. The output of each successful node can view. With these features, the user can clearly understand the analysis procedure of the data, the process interruption and the cause of the interruption. It also help the user to solve the problem and continue the previous data analysis operation.

Besides, the parallel platform provides a real-time monitoring part of the task, which can be used to view the running status of the current submitted data analysis process and know the processed component node and its state of the current analysis process, to assist the user.

In addition to the basic data mining functions, the parallel platform also provides various types of functional operations, such as scheduling flow, text mining, social network analysis, deep learning, and web reporting, for users to process data analysis in a multi-dimensional perspective.

V. CONCLUSION

For the demand of power big data business analysis and mining along with the development of smart grid, this paper designs and develops a distributed parallel data mining platform for power big data. The parallel platform adopts a highly reusable distributed framework and designs a scalable parallel algorithm library. It integrates nearly 100 well-running performances and highly parallelized algorithms, which are partially superior to existing open source algorithm libraries MLib and Hive tools. A variety of deep neural network, machine learning, statistical analysis and other categories of parallel general mining analysis algorithms and special algorithms for power data analysis business needs are involved in library. In addition, the graphical interactive interface Studio provided by the parallel platform indicates the data flow direction and dependence relationship with the workflow diagram. It also visually displays the system analysis process, the current execution status and the analysis result. Then the Studio supports the user to customize the data analysis tasks by clicking and dragging without the operations to write programs, which reduces the threshold for users to perform big data analysis tasks.

REFERENCES

[1] Holmes G, Donkin A, Witten I H. Weka: A machine learning workbench[C]//Intelligent Information Systems, 1994. Proceedings of the 1994 Second Australian and New Zealand Conference on. IEEE, 1994: 357- 361.
[2] Hofmann, M., Klinkenberg, R. RapidMiner: Data mining use cases and business analytics applications[M]. CRC Press, 2013.
[3] Berthold M R, Cebron N, Dill F. KNIME: The Konstanz Information Miner[J]. Acm Sigkdd Explorations Newsletter, 2006, 11: 26-31.
[4] An ZHUO. Research and Implementation of Big Data Analysis Platform Based on P2P Scalable Architecture[D]. Tsinghua University, Beijing,China, 2012.
[5] Yu L, Zheng J, Shen W C, et al. BC-PDM: data mining, social network analysis and text mining system based on cloud Computing [C]//Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2012: 1496-1499.
[6] De Francisci Morales G. SAMOA: A platform for mining big data streams[C]//Proceedings of the 22nd International Conference on World Wide Web. ACM, 2013: 777-778.
[7] Qing HE, Fuzhen ZHUANG, Li ZENG. PDMiner: Parallel Distributed Data Mining Platform Based on Cloud Computing[J]. Science China, 2014, 44: 855-871.
[8] Li W, Cheng H L, Peng Y, et al. Visualized data mining platform based on the Spark[J]. Chinese Association of Automation System SimulationProfessional Committee, 2014.
[9] Jun LEI, Hangjun YE, Zesheng WU, Big-Data Platform Based on Open Source Ecosystem[J]. Journal of Computer Research and Development, 2017, 54: 80-93.
[10] Guo T, Xu J, Yan X, et al. Ease the Process of Machine Learning with Dataflow[C]//Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. ACM, 2016: 2437-2440.
[11] Bu Y, Wu B, Chen Y. BDAP: A data mining platform based on Spark[J]. Journal of University of Science Technology of China, 2017, 47: 358-368.
[12] Yu K. Large-scale deep learning at baidu[C]//Proceedings of the 22nd ACM international conference on Information and Knowledge Management. ACM, 2013: 2211-2212.
[13] Abadi M, Barham P, Chen J, et al. Tensorflow: a system for large-scale machine learning[C]//OSDI. 2016, 16: 265-283.

