ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 2 - 2026

Research on the Collaborative Distributed Computing Architecture of Power Marketing Headquarters-Provincial Companies for Massive Heterogeneous Data

Dongze Wang[1,*], *Lixuan Gao*[1], *Liang Gao*[1], *Yuehui Han*[1] *and Kun Gan*[1]

[1] State Grid Co., Ltd. Customer Service Center, Tianjin, 300000, China

SUMMARY: *Aiming at the potentially associated multi-group features and high-dimensional multi-source heterogeneous data in the power grid, this paper utilizes distributed technology to address data updating between multi-level heterogeneous spatial databases. The tasks of the centers at all levels are to receive, store, process, manage and apply various types of resource data needed for resource management and social services, and to realize synchronous or asynchronous updates with remote data from higher-level data centers and lower-level data centers through the Resource Information Network (RIN). Through the distributed parallel framework and deployment mapping and subsumption services, the analysis and calculation process of power marketing data is established, and the data processing workload of the headquarter-provincial company is assigned to the high-performance computing cluster for execution, and its calculation results are stored in a distributed manner so as to facilitate redistributed calculations after changing the algorithms and parameter settings of power marketing data. The sparse modeling realized by LASSO regression can effectively reduce the computation time, and can support the steady progress of the online assessment of voltage stability margin. The experimental results show that the load forecasting error rate of this paper's architecture is 3.5%, the accuracy of customer segmentation is 91%, and the average acceleration ratio is 3.83, and the framework can provide guidance and support for the stable operation of the power system.*

KEYWORDS: *multi-source heterogeneous data; distributed technology; data update; distributed parallel framework; power marketing data*

1 Introduction

With the accelerated pace of informatization of the national power grid, the amount of data in electric power companies shows an exponential growth [1]. Multi-source heterogeneous data is the main component of the data of power companies. The acquisition of multi-source heterogeneous data by electric power companies mainly comes from a diversity of sensors, including electronic equipment, grid terminals, transmission lines, transmission lines, power production and transmission and sale of electricity, and the frequency of real-time collection data. In addition, power regulation and trading of power resources will also generate more information interaction data records [2-4]. Usually, the challenges posed by massive heterogeneous data in electric utilities include several aspects such as storage, transmission and information processing [5]. In the face of a large amount of heterogeneous data from multiple sources, how to store, call, compute and manage them effectively has become an

*Sg_wdz@126.com
https://doi.org/10.65102/is2026651
Published: 30 April, 2026

urgent problem.

Fracas, P et al. proposed a techno-economic model of two interconnected hybrid microgrids (MGs) whose power and heat dispatch strategies are managed using sequential least squares planning (SLSQP) optimization techniques.The MG combines multiple thermoelectric power generation, transmission, and distribution systems into a single whole that tightly integrates weather-dependent distributed renewable energy generators with multiple stochastic load profiles [6]. Ding, T et al. introduce the basic concepts and EH modeling methodology and provide a systematic review of optimization methods for EH planning, operation and trading as well as algorithms for state-of-the-art solutions. An Internet of Things (IoT)-based EH control structure is analyzed and the corresponding state estimation, communication, and control methods for managing large EH datasets are reviewed [7]. Keskin, N. B et al. Considering an electric utility serving retail electric customers on a discrete time horizon, they design a data-driven joint spectral clustering and feature-based pricing strategy and show that their strategy achieves near-optimal performance [8]. Wang, B et al. in order to improve the accuracy and fairness of IoT data sharing, fully consider the heterogeneity of the participants and improve the data valuation and profit distribution in IoT data sharing based on electricity retailing. Data valuation should be related to the attributes of IoT data purchasers, where risk appetite of electricity retailers is selected as a characteristic attribute, and data premium rate is proposed to characterize its impact. Profit allocation should fairly measure the marginal profit shares of power retailers and data brokers (DBs), thus an asymmetric Nash bargaining model is used to ensure that they receive a reasonable profit based on their contributions to the IoT data sharing consortium. Taking into account the heterogeneity of the participants, the proposed IoT data sharing is suitable for large consortiums exchanging IoT data with multiple electricity retailers and DBs [9]. Ma, Y et al. utilize a shared Battery Energy Storage System (BESS) to take on the PFR obligations for multiple wind and PV power plants, while providing commercial Automatic Generation Control (AGC) services in the ancillary services market. An optimal bidding strategy in the pre-hourly energy and ancillary services markets is proposed to maximize the benefits of the BESS. The optimization model is constructed as an opportunity-constrained optimization problem, which is transformed into mixed-integer quadratically constrained programming using a distributionally robust optimization technique based on the Wasserstein metric [10].

Levin, T et al. identified challenges associated with improving modeling capabilities to inform decarbonization policies and power system investments and improve societal outcomes throughout the clean energy transition through electric energy storage in capacity expansion modeling [11]. Lu, X et al. developed an integrated model to assess solar PV potential and its cost competitiveness from 2020 to 2060 by considering multiple spatial and temporal factors. The cost competitiveness of solar energy was found to allow pairing with storage capacity to provide 7.2 PWh of grid-compatible electricity to meet 43.2% of China's demand in 2060 at less than 2.5 cents/kWh [12]. Cheng, Z et al. investigated the cybersecurity of distributed AC optimal currents (ACOPF) against False Data Injection Attack (FDIA). A collaborative distributed ACOPF solver based on the concept of dyadic decomposition was constructed, and based on this, a theoretical framework was developed for modeling distributed ACOPF and its cybersecurity in the presence of FDIA. Within this proposed cybersecurity framework, a reputation-based peer-to-peer trust management system (TMS) is proposed to ensure system resilience to FDIA. Finally, the proposed TMS is validated on the IEEE 69 bus benchmark system [13]. In order to overcome the limitations of analysis algorithms due to implementation in MapReduce programming model, Sun et al. discussed the problems when analyzing big data, inefficiencies, memory constraints that turn into algorithmic constraints, the need to overcome such problems requires the development of new

non-MapReduce distributed computational frameworks [14]. Parallel and distributed computing systems are key to support large-scale applications such as scientific simulations, big data and AI, as explored by Perera et al. As system complexity increases, optimizing performance faces many challenges.

Electric power companies usually exist in the form of large groups, and collaborative management of power data through the headquarters-provincial companies, the headquarters is responsible for the development of marketing plans, industry policy at the same time need to adjust the distributed resources, data analysis and decision-making. Provincial companies are responsible for the policies, strategies, plans, and decision-making results of the head office to implement the operation of transcription of electricity, business access, risk supervision, and load management. Collaborative management between the head office and provincial companies helps to improve data processing efficiency, increase computing power, and reduce unnecessary data transmission. Therefore, this paper designs a dimension model, a fact model, an aggregation model, and an application model based on the marketing data supply system and the supply center model of the enterprise-level data middleware, and establishes a distributed storage topology consisting of a server client, metadata server clusters, and intelligent storage clusters. The distributed storage topology composed of server clients, metadata server clusters and intelligent storage clusters is established. Design real-time data supply, offline data supply and other link channels, which can respond to system requirements in time. Calculate Map-Reduce time sequence diagram, map and merge service for massive heterogeneous data information of electric power, and realize real-time data update. Adopting adopting multivariate cooperative learning algorithm and calculating loss function to improve the power data generalization ability. Realize the storage and access of massive heterogeneous data through power marketing distributed. The design of Map-Reduce timing diagram of headquarters-province cooperative computing is used to realize power marketing data mapping and merging to achieve the effect of parallel processing of tasks. Utilizing multivariate cooperative learning algorithms for effective data fusion.

2 Data supply system design

2.1 System framework

Heterogeneous data distributed architecture is shown in Fig. 1, the heterogeneous distributed data center adopts a mesh linkage, any 2 data centers can be connected to each other, but in reality there is generally a primary and secondary relationship between the data centers, so the connection structure of the data center is expressed as a tree hierarchy [15].

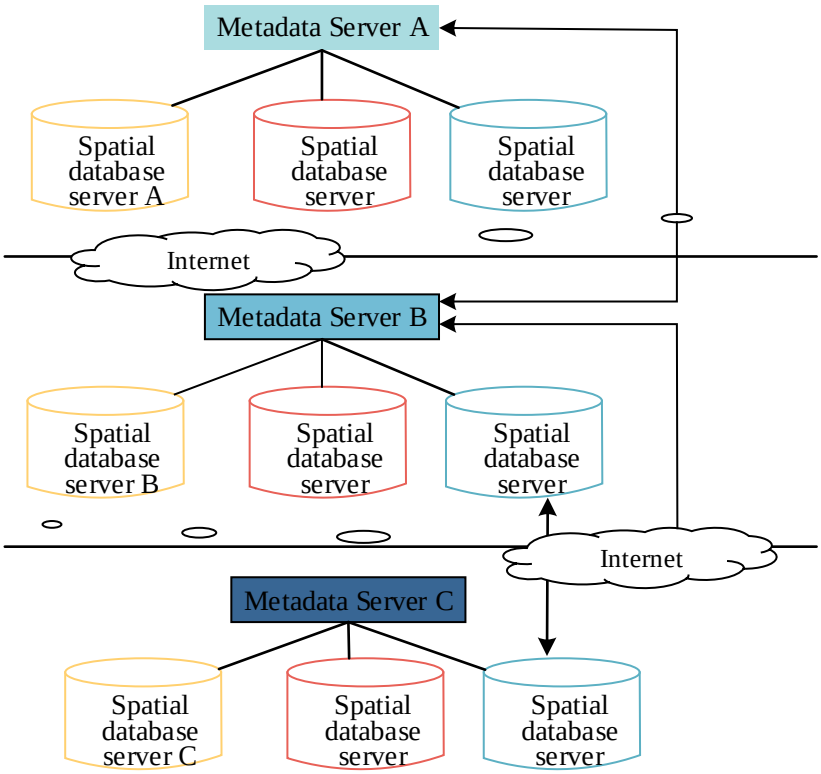

Figure 1: Heterogeneous Distributed Data Center Architecture

Since the data needs to be transmitted remotely and the data update and synchronization is mainly carried out through the private network, it mainly relies on Web Service technology to realize data service. The general principle is that where there is data where there is service, in Figure 1, A data center has a spatial database server A, it is necessary to provide A database access to the Web service. Similarly, B data center needs to provide Web services for B database access. But the Web service is a stateless application, only responsible for providing services, so there is no relationship between the A data center database service and the B data center database service, you need to coordinate the communication through the metadata server to establish a logical connection between the two, and then realize the data update.

2.2 Data processing core

For large-scale power grids, the database cannot cover all operating conditions at the initial stage, and needs to be continuously updated and progressively improved, especially considering the large amount of heterogeneous data in the new power system and the high penetration of new energy sources, a high-quality database autocorrelation program is of great significance to improve the coverage of the database and to assess the accuracy of the model [16].

For the processing of power marketing data, in order to adapt to the change of system information, a local root-mean-square indicator $\gamma_{locRMSE}$ is used to measure the data completeness and accuracy, whose defining formula is as follows:

$$\gamma_{locRMSE} = \left[\left[M' - X'\beta \right]^T W(Z_0) \left[M' - X'\beta \right] \right]^{1/2} \quad (1)$$

Among them, $\gamma_{locRMSE}$ represents the local root mean square index, which is used as a basis for online judgment of the degree of outliers of the running point, and then triggers the database update. M', X' denote the target data matrix after clustering process, X' is the independent variable matrix, and $W(Z_0)$ is the Gaussian kernel function, respectively.

2.3 Two-tier data provisioning architecture and technology based on enterprise-level data middleware

The study designs an efficient, stable and scalable supply architecture based on the enterprise data middle office, and provides comprehensive theoretical and practical support for enterprises through the design of the marketing data supply system and the design of the supply center model for the enterprise-level data middle office.

Table 1 shows the data architecture of the power marketing headquarters and the provincial level, analyzing the marketing common data supply demand, the data supply thematic model library and the two-level supply capacity, sorting out the marketing common data supply demand by analyzing factors such as the marketing business, market demand, operation and the external environment, analyzing the priority and key areas of data supply by taking into account the business demand and the assessment of the capacity of the data supply center, and prioritizing the protection of the core demand data supply.

High-quality supply of data elements is the source of releasing the value of data application. Promoting the construction of the two-level data supply system and promoting data integration, interoperability and interoperability effectively reduces the cost of data circulation, improves the efficiency of data supply, provides basic support for high-quality supply of data, strengthens the Company's competitiveness, enhances the industry position and lays a solid foundation for the Company's sustainable development. In addition, the promotion of the results also helps to promote the development and application of related technologies, promote the digital transformation of the entire industry, and promote the widespread application of data-driven and intelligent applications in the energy sector [17, 18]. The analysis of marketing common data supply demand and two-level supply capacity can promote cooperation and synergy between different business sectors.

Table 1: Power Marketing Headquarters and Provincial Data Architecture

Data application classification	Report dashboard class		Special analysis category	Ad hoc query class	
Two level data supply demand	Non real time business detail data	Bill data	Power outage information data	Analysis of Abnormal Fluctuations in Indicators	Real time business detail data
Bata domain	Analysis Domain (Data Platform)	Analysis Domain (Data Platform)	Analysis Domain (Data Platform)	Processing Domain (Business Center)	AnalysisDomain (Data Platform)
Two level data supply technology roadmap	MaxCompute/ DWS-> SG-UEP-> MaxCompute	MaxCompute/ DWS-> SG-UEP-> MaxCompute	DataHub/kafka-> SG-UEP-> DataHub+Flink-> ADB/RDS	MaxCompute/ DWS-> SG-UEP-> MaxCompute	DataHub/kafka-> SG-UEP-> DataHub+Flink-> ADB/RDS

2.3.1 Model library design

In order to improve the efficiency of data supply for marketing professional scenarios, for the actual needs of marketing professional scenarios, combing the common needs involves basic data and result data, and carrying out the design of the data supply model library around the four aspects of dimensional model, factual model, aggregation model and application model [19]. Specifically as follows:

(1) Dimension model design through the model definition and naming guidance, to promote the marketing common data set dimension table model design work to refine the dimensions corresponding to the business requirements, around the user attributes, type of electricity consumption, power supply code, registration time and other specific descriptive information, through the determination of the main dimensional table, determine the relevant dimension table, determine the dimension attributes of the three steps to determine the dimension model table structure.

(2) Fact model design According to the needs of common dataset construction, select business applications to carry out common dataset fact table model design work. Around the business application scenarios, based on the scenario theme, data change type, alteration time, transmission mode, source location and other summarization needs, the fact statement additional field information, and to ensure that the facts and dimensions of the granularity of the facts to maintain consistency, to generate the fact table model.

(3) Summary model design is used to store summary data with more reserved dimensions, and with the business scenario requirements as the modeling driver, different levels of attribute degradation are carried out for the physical model of the factual detail table to form a summary wide table model combining dimensions and facts, and to improve the efficiency of data analysis and query speed.

(4) The application model is directly oriented to the application of business requirements, through the common data requirements of business topics involved in the expansion, metering and customer service scenarios, convergence of business data attributes, selection of associated fields, based on the further generalization and integration of the summary wide table, increase the business scenarios, operational flags, effective flag field information, to generate the application results table model [20].

Combined with the headquarters side and the actual business scenarios at the provincial level, in the face of data such as report Kanban, business monitoring, thematic analysis, and on-the-spot query, the data supply channels such as real-time links and offline links are formulated based on the data center, and data supply guarantee is provided to ensure that the data is correct and available.

2.3.2 Link channel design

Design link channels for real-time data supply and offline data supply for the characteristics of two-level application of marketing data as follows:

(1) The data table in Marketing 2.0 is synchronized with log collection through the OGG/DRS tool, the data in Datahub/Kafka is archived to Maxompute for full storage and batch analysis and calculation, Flink subscribes to the Topic in Datahub and associates the dimensional data for processing and calculation. Use Flink to write data to the calculation results in real time, Flink's write operation needs to rely on the business primary key of each piece of data in order to realize fast by partition insert, update, delete and other operations. On-demand business detail data is written directly to the calculation result library through Datahub, and the two-level detail data transmission is achieved by subscribing to Datahub/Kafka through SG-UEP, and the synchronization of the two-level dimension data is achieved through SG-UEP offline.

(2) The non-real-time business detail data generated by Marketing 2.0 is written to the posting source layer of the data middle station through OGG+Datahub/Kafka, the wide table is formed through the sharing layer, and the resultant data is imported to the analysis layer after the analysis layer calculates and summarizes the resultant data, and the report and index data query service is provided to the outside world, and the data coherence of the two levels of the sharing layer and the analysis layer is realized through SG-UEP.

(2) The two-level data supply application guarantee is centered on credible source, stable link, and cooperative operation, and is oriented to end-to-end full-link and full-state monitoring and analysis of the two-level middle station, so as to build a full-link monitoring application for the two-level data supply of marketing, and to guarantee the transparency of the information of the two-level link and ensure credible source of data from the management side. From the technical side, it guarantees the stable operation of the two-level link, realizes timely detection of faults, and ensures the stability and reliability of the data in the middle station. From the operation side, it collaborates with the two levels to support the middle station to dispose of alarms, respond to business needs, realize one-stop service application, and guarantee stable and efficient operation and credible and available data.

3 Collaborative Distributed Computing Architecture Design for Massive Heterogeneous Data

3.1 Distributed Storage Topology

For the current mainstream local or distributed file systems, there is no file system suitable for electric power massive heterogeneous data storage system, which can not simultaneously meet the requirements of stability, centralized management of massive files, high disk utilization for massive pictures and small files, ability to perform rapid retrieval and playback of massive video files according to the storage time, and separation from metadata [21, 22].

In order to solve the above problems, a fully distributed block file system suitable for power massive heterogeneous data is designed and implemented. This block file system architecture can either be used as a file system for NAS, combined with NAS for storage, or use iSCSI for direct block file network storage.

This system is composed of three major parts: server client, metadata server cluster, and intelligent storage cluster, and the distributed storage topology of power marketing data is shown in Figure 2. The metadata cluster consists of more than two dedicated servers that manage the metadata of the file system in the data center, including file directory tree organization, attribute maintenance, file operation log records, authorized access, retrieval mapping, etc. It manages the namespace of the entire storage system and provides a single system image to the server client. Intelligent storage cluster consists of more than 3 dedicated storage servers, DataNode nodes), which are the units that actually store the data and are the storage resource providers of UniMAS. The server client is installed on the application server where the user needs to use the storage resources, and provides a standard POSIX interface to read and write access to the storage resources of UniMAS.All metadata servers of UniMAS are online at the same time to provide services, and every two metadata servers are in a group, and each group of metadata servers will be backed up by each other internally, to ensure that within a group, if any of them has a problem, the other one will have a problem with the metadata. If any one of the metadata servers in a group has a problem, the other metadata server will have the complete data of this group and take over the service without interrupting the business.

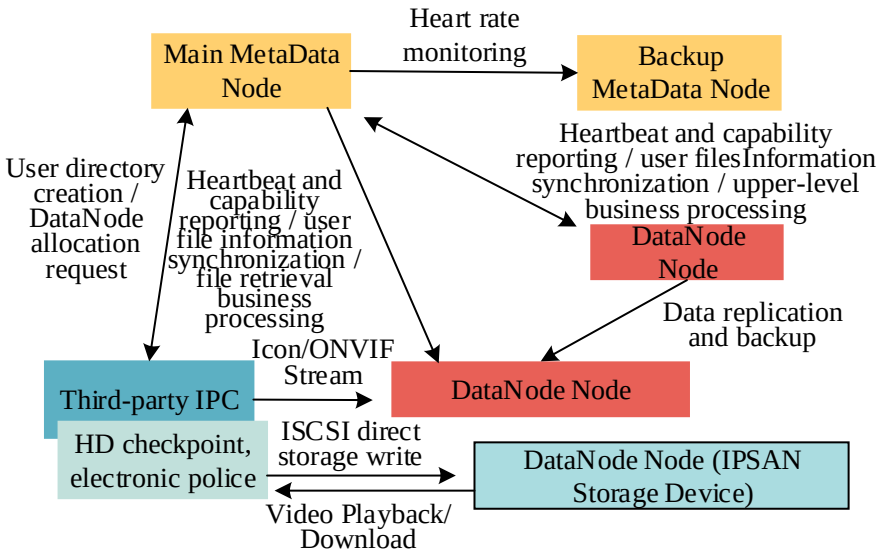

Figure 2: Topology diagram of distributed storage for power marketing data

3.2 Headquarters-province co-calculation

The headquarters-province co-computing layer passes attribute computation parameters to the mapping service and the subsumption service, reads a slice of data from a distributed file data block according to the sub-block data body file size, and converts it to a set of key-value pairs, obtains all the computation units, and extracts the distributed power marketing sub-block data. Then, the mapping task traverses the data in each sub-block and performs the specified operation on each key-value pair to compute by key value. Wherein, key-value pair refers to a pair of structures composed of heterogeneous data bodies of power marketing, corresponding to its business scope, and after the completion of the mapping task, its output intermediate results will be distributed to different subsumption tasks based on the hash value of the key, and in this process, the values of the same key will also be sorted and merged so that the Reduce task can process them efficiently. Then, asynchronous messages activate the merging service to perform the merging work after confirming that all mapping services have completed the intermediate analysis results. Each merging task receives the list of intermediate results assigned to it and performs the specified operations according to the keys, traverses and merges the final data block files to form the distributed file results, and finally generates the final outputs, which are uploaded to the sub-block storage service node to complete the distributed mapping. Similarly, the process of merging is built in the framework of executing multi-threaded parallelism.

Figure 3 shows the headquarters-province collaborative computing Map-Reduce timing, through the design of the headquarters-province collaborative computing Map-Reduce timing diagram, it is able to carry out mapping and merging services for the massive heterogeneous data information of electric power, and realize parallel processing of tasks.

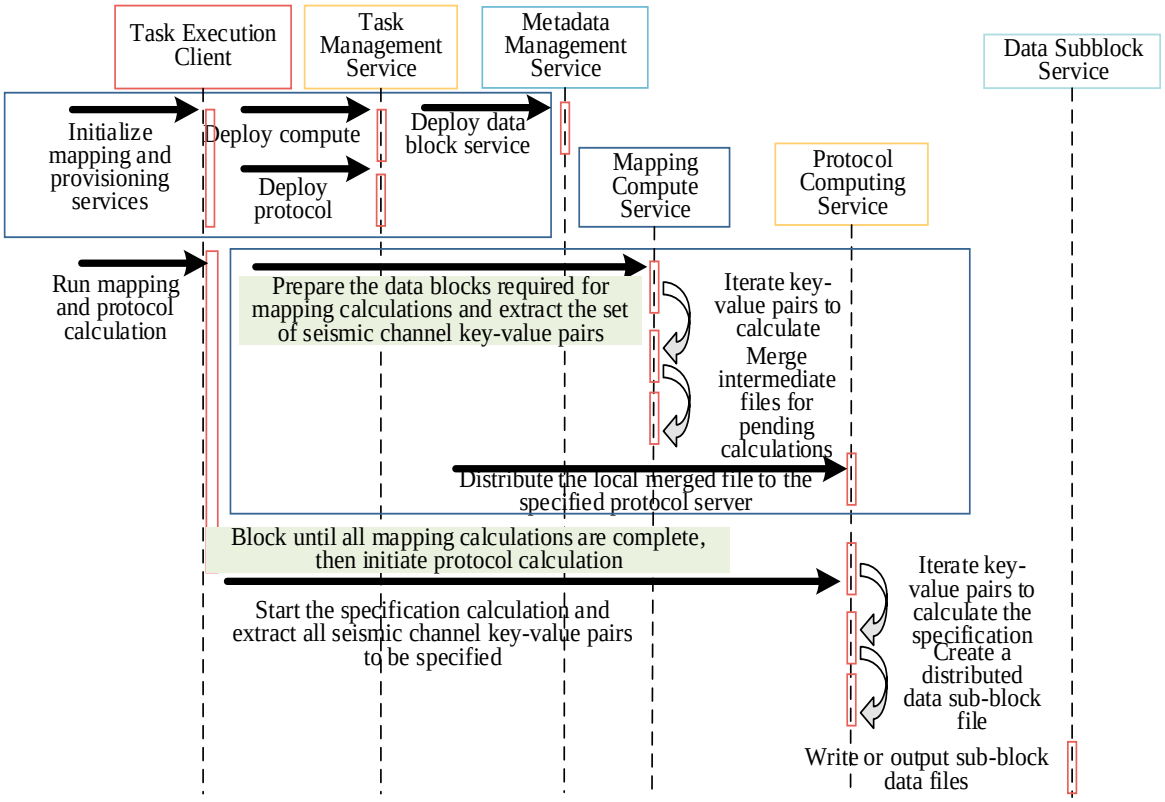

Figure 3: Headquarters Province Collaborative Computing Map Reduce Time Series

3.3 Loss function calculation

In this paper, a multivariate cooperative learning algorithm is used to obtain the correlation between heterogeneous data from multiple sources, so as to realize the deep fusion of data. The regression model is fitted by optimizing the loss function, and the loss function expression for LASSO regression is:

$$\min_{\beta}\left\{\frac{1}{2}\left[M^{'}-X^{'}\beta\right]^{T}W\left(Z_{0}\right)\left[M^{'}-X^{'}\beta\right]+\lambda\left|\beta\right|\right\} \tag{2}$$

where λ denotes the regularization factor and β denotes the matrix of regression fit coefficients.

Weighted during the fitting process, the objective function selects the L1 regularization penalty term. As λ increases, some of the features represented by the regression fit coefficients may decrease to 0, which means that the L1 paradigm tends to compress coefficients with low potential predictive value, thus generating sparse coefficients. Therefore, it can be said that the LASSO regression eliminates ineffective features without loss of accuracy and thus simplifies the extracted features even further, playing a kind of "parameter selection" [23, 24]. Sparsity means that very few entries in the matrix are non-zero, and the L1 paradigm has the property of generating many coefficients with zero values or very small values with few large coefficients.The emergence of the L1 paradigm results in a loss function that is not always derivable, however, the sparsity of the solution of the L1 paradigm makes it possible to use it in conjunction with sparse algorithms, such as coordinate descent, least angle regression, and so forth, which makes the computation more efficient. Since LASSO

regression is more robust and generalizable than traditional machine learning techniques, it can be applied in voltage stability margin assessment work to effectively deal with the cases of missing data, excessive dimensionality, continuous or scattered data.

3.4 Performance indicators

In order to realize the comparison of model performance across datasets, this paper adopts residual squared error R2, root mean square error γ_{RMSE} as the comparison index:

$$R^2 = 1 - \sum_{i=1}^{n}\left(M_i - \hat{M}_i\right)^2 / \sum_{i=1}^{n}\left(M_i - \bar{M}_i\right)^2 \tag{3}$$

$$\gamma_{\mathrm{RMSE}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[M_i - \hat{M}_i\right]^2} \tag{4}$$

where M_i represents the actual business data and $\hat{M}_i$ represents the fused data $\bar{M}_i$ is the mean of M_i.

After training, the model will be tested for performance on an unseen test set, which puts requirements on the generalization performance of the model. In grid operation, the operating points are probabilistically distributed around the equilibrium point, and a qualified evaluation model must have good generalization performance.

4 Distributed Computing Processing Verification and Analysis

4.1 Experimental environment construction

4.1.1 Hardware environment

Spatial data are stored in the database server. Provincial data centers, municipal data centers and county data centers store data of the whole province, municipal data and county data respectively, so it can be seen that the data volume of the provincial data centers is very large, and the data of the data centers will have to be provided for the use of the government system and social services in the future, so it puts forward high requirements for the performance and security of the database servers.

Database server configuration as shown in Table 2, the database server needs at least 2 sets of RAID5 to ensure the security of the database server operating system. At the same time, disk arrays are needed to realize dual-machine hot standby to ensure data security, and the minimum capacity of disk arrays should not be less than 2 TB. The database server is connected to a Fibre Channel switch and a tape library is connected to the database server for data backup at the same time. The operating system of the database server can be Windows 2003, Linux or UNIX, and the commercial database is one of SQLServer and Oracle.

The metadata server stores subscription, release and update information of data center data at all levels. The operating system Windows2000, the database can be SQLServer or Oracle. WebService server is the server used to publish WebService service for spatial data updating and services for data publishing subscription of metadata server at all levels. Installation of Windows 2003, IIS6.0. Server CPU requires IntelXeonMP3.0GHz/4MB or more, and RAM

requires 2G or more. WebService server can also be placed on the same server as the metadata server.

There are 2 kinds of network connections between data centers at all levels, one is private network and the other is Internet. The department of land and resources usually establishes an internal private network or rents lines from telecommunication carriers to form a private network with a good network connection and a bandwidth of more than 2M. If it is through the Internet, the following methods are used to achieve the reliability of the system, a separate metadata server is set up at each level of the data center, and the data updates are transmitted through GML, or in the case of whole file updates, in chunks, to minimize the amount of data transmitted each time and to ensure high reliability, with the ultimate goal of achieving an element-level update, in which only the elements that have been updated are transmitted.

Table 2: Database Server Configuration

Item	Specification Requirements	Configuration Requirements
CPU	Supports Inter Xeon MP 3.0 GHz/4 MB or higher CPU	4 units of Inter Xeon MP CPU with 3.0 GHz/4 MB or higher (minimum 4 units)
Chipset	ServerWorks GC-HF supports 8-way interleaved memory access	-
Memory	ECC DDR memory, supports Chip Kill technology, maximum support 32 GB, supports memory mirroring and memory hot backup unit	8 GB
PCI	No less than 7 hot-swappable 100 MHz PCI-X, no less than 1 PCI32 33 MHz	-
Internal Hard Drive	Supports 36 GB, 72 GB, and 146 GB Ultra320 SCSI hard drives, supports no less than 5 hot-swappable hard drive slots, supports 2+3 separated SCSI backplanes, rotation speed ≥10 KRPM	At least 3 pieces of 73 GB hard drives, configured as RAID5
RAID Controller	Ultra320 RAID controller, no less than 128 MB with battery-backed high-speed cache, no less than 2 channels	1 RAID controller
Network Card	No less than 4 100/1000BASE-T Ethernet network cards	-
Hot-swappable Redundant Power Supply	No less than 2 redundant hot-swappable power supplies	2 redundant power supplies

4.1.2 Software environment

The simulations are implemented on Matlab 2014a, and all the simulations are obtained by running on a personal computer with Intel(R) Core(TM) i5-4210U CPU@1.70GHZ 2.39GHZ and 8G of RAM. The experimental data are derived from real marketing data of a provincial power company, including 100GB of structured electricity consumption data, 50GB of semi-structured business data, and 80GB of unstructured voice image data, and the massive simulation data are generated by the data expansion tool, with a total data volume of 500GB.

4.2 Experimental results and analysis

4.2.1 Co-computing performance analysis

Comparing the performance of this paper's architecture and the traditional architecture under the same file size, Table 3 shows the running time of the traditional architecture, and the average acceleration ratio of the traditional architecture under different capacity data is 3.87. The system throughput gradually rises from 42.55MB/s to 80.56MB/s, and the performance of parallel computing is poor. When the file size is 320/MB, the communication rate of the traditional architecture is close to decreasing gradually to 60.12/Mbit·s^{-1}, which can not meet the needs of coordinated computation cooperation between the power marketing headquarters and the provincial companies under massive heterogeneous data.

Table 3: Runtime of Traditional Architecture

File size/MB	Serial time/s	Parallel time/s	Speedup ratio	Throughput MB/s	Transmission rate /Mbit·s^{-1}
10	30.82	7.84	3.93	42.55	98.50
20	50.24	12.64	3.97	49.84	90.42
40	98.62	25.67	3.84	54.65	82.10
80	196.55	51.11	3.85	61.85	75.30
160	390.23	102.67	3.80	70.23	69.23
320	786.55	206.76	3.80	80.56	60.12

Distributed computing architecture running time is shown in Table 4, the average speedup of this paper's architecture is 3.83, which is due to the security and reliability of this paper's architecture benefit distributed framework in heterogeneous data transfer. When the file size is 320/MB, the throughput of this paper's distributed computing architecture is 20.40MB/s, which is a small variation and has significant advantages of parallel computing. The transmission rate is reduced from 98.91/Mbit·s^{-1} to 93.561/Mbit·s^{-1}, which is still able to better meet the actual needs of the current power marketing application platform.

Table 4: Distributed computing architecture runtime

File size/MB	Serial time/s	Parallel time/s	Speedup ratio	Throughput MB/s	Transmission rate /Mbit·s^{-1}
10	29.87	7.65	3.90	14.64	98.91
20	48.76	12.59	3.87	17.35	97.20
40	96.04	25.26	3.80	18.20	96.33
80	195.99	51.58	3.80	19.00	95.05
160	388.62	102.56	3.79	19.73	94.24
320	782.42	206.21	3.79	20.40	93.56

Compared with the traditional architecture, for the same file, the running time is reduced. Compared with the serial traditional distributed architecture algorithm, after the improvement of parallel optimization, the running time is reduced to about 1/3 of the original. two comparisons can be concluded that, compared with the traditional distributed architecture, the architecture proposed in this paper can greatly improve the encryption efficiency and reduce the performance requirements of the computers, so as to provide some theoretical support and ideas for the problem that the computer performance can not be processed in time due to the continuous expansion of the data scale. Certain theoretical support and ideas.

4.2.2 Analysis of business application effects

The results of voltage magnitude state estimation using the traditional architecture and the method proposed in this paper are compared and analyzed, and the comparison of business application effects is shown in Table 5. In the load prediction task, the prediction based on the processed data of this architecture has a prediction error rate of 3.5%, which is lower than that of 7.8% in the traditional architecture. In the customer segmentation task, the customer segmentation accuracy rate of this architecture reaches 91%, which is higher than the 83% of the traditional architecture, verifying the application value of this architecture in the actual business.

Table 5: Cmparison of Business Application Effects

Business task	Error rate/accuracy of this architecture (%)	Traditional architecture error rate/accuracy (%)
Load forecasting	3.5 (error rate)	7.8 (error rate)
Customer segmentation	91 (accuracy)	83 (accuracy)

4.2.3 Comparison of load results

The error distribution of the load forecasting results is shown in Fig. 4, where the plotting logic of the computational errors using the traditional set architecture and the method proposed in this paper visualizes the error advantage of this architecture in the load forecasting task. The traditional architecture has huge error fluctuations in load forecasting, with the highest fluctuation reaching 1.2. While the error of this paper's architecture is very smooth and always stays within 0.2. It shows that the architecture of this paper has high efficiency and is more accurate in predicting the power marketing load data.

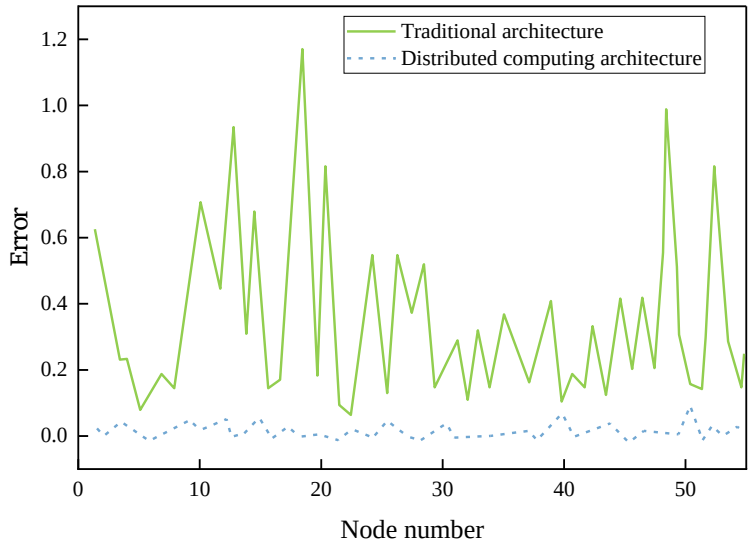

Figure 4: Comparison of Load Forecasting Error Distribution

Experimental results show that this architecture is better than the traditional architecture in terms of heterogeneous data fusion efficiency, collaborative computing performance and business application effect, and it can meet the demand of power marketing headquarters-provincial companies for collaborative processing of massive heterogeneous data, and provide effective technical support for lean management of power marketing.

5 Conclusion

This paper proposes a research on collaborative distributed computing architecture for massive heterogeneous data for power marketing headquarters-provincial companies. With the help of advanced distributed architecture and intelligent algorithms, it breaks through the technical bottlenecks of traditional marketing systems and realizes the intelligence, refinement and personalization of marketing business. Experimental results show that the average acceleration ratio of this paper's architecture under different capacity data is 3.83, the error rate of load forecasting in business applications is 3.5%, and the accuracy rate of customer segmentation is 91%, which makes the platform basically satisfy the current requirements of electric power big data analysis and big data application development. However, through research and development, it is also found that the current platform has the following deficiencies, which require further optimization of heterogeneous data fusion algorithms in subsequent research and development, combined with deep learning technology to improve the feature extraction accuracy of unstructured data, and adapt to the growing scale and type of data.

About the Author

Dongze Wang earned his Master's degree from the School of Control Science and Engineering at Hebei University of Technology. His work primarily focuses on the digital construction of power marketing and data application analysis.

Lixuan Gao, born in Zhangye City, Gansu Province in 1992, graduated from Lanzhou University of Technology majoring in electronics and communication engineering with a master's degree. The main research direction is cryptography.

Liang Gao, Born in Anshan City, Liaoning Province, China in 1980, he graduated from Northeast Electric Power University with a master's degree. The main research direction is the power system and its automation.

Yuehui Han, Born in Tangshan City, Hebei Province in 1995, China, he graduated from the School of Computer Science of Northwest University of Technology with a master's degree. The main research direction is semiconductor device simulation.

Kun Gan, born in Xinyang City, Henan Province in 1987, graduated from Jiangsu University with a bachelor's degree in information management and information systems.

References

[1] Wang, Y., Jia, M., Gao, N., Von Krannichfeldt, L., Sun, M., & Hug, G. (2022). Federated clustering for electricity consumption pattern extraction. *IEEE Transactions on Smart Grid*, *13*(3), 2425-2439.

[2] Imani, M. H., Bompard, E., Colella, P., & Huang, T. (2025). Data analytics in the electricity market: a systematic literature review. *Energy Systems*, *16*(1), 1-35.

[3] Bjarghov, S., Löschenbrand, M., Saif, A. I., Pedrero, R. A., Pfeiffer, C., Khadem, S. K., ... & Farahmand, H. (2021). Developments and challenges in local electricity markets: A comprehensive review. *IEEE Access*, *9*, 58910-58943.

[4] Hogan, W. W. (2022). Electricity market design and zero-marginal cost generation.

Current Sustainable/Renewable Energy Reports, *9*(1), 15-26.

[5] Solyali, D., Safaei, B., Zargar, O., & Aytac, G. (2022). A comprehensive state-of-the-art review of electrochemical battery storage systems for power grids. *International Journal of Energy Research*, *46*(13), 17786-17812.

[6] Fracas, P., Camarda, K. V., & Zondervan, E. (2023). Shaping the future energy markets with hybrid multimicrogrids by sequential least squares programming. *Physical Sciences Reviews*, *8*(1), 121-156.

[7] Ding, T., Jia, W., Shahidehpour, M., Han, O., Sun, Y., & Zhang, Z. (2022). Review of optimization methods for energy hub planning, operation, trading, and control. *IEEE Transactions on Sustainable Energy*, *13*(3), 1802-1818.

[8] Keskin, N. B., Li, Y., & Sunar, N. (2025). Data-driven clustering and feature-based retail electricity pricing with smart meters. *Operations Research*, *73*(5), 2636-2660.

[9] Wang, B., Guo, Q., Xia, T., Li, Q., Liu, D., & Zhao, F. (2023). Cooperative IoT data sharing with heterogeneity of participants based on electricity retail. *IEEE Internet of Things Journal*, *11*(3), 4956-4970.

[10] Ma, Y., Hu, Z., & Song, Y. (2022). Hour-ahead optimization strategy for shared energy storage of renewable energy power stations to provide frequency regulation service. *IEEE Transactions on Sustainable Energy*, *13*(4), 2331-2342.

[11] Levin, T., Bistline, J., Sioshansi, R., Cole, W. J., Kwon, J., Burger, S. P., ... & Botterud, A. (2023). Energy storage solutions to decarbonize electricity through enhanced capacity expansion modelling. *Nature Energy*, 8(11), 1199-1208.

[12] Lu, X., Chen, S., Nielsen, C. P., Zhang, C., Li, J., Xu, H., ... & Hao, J. (2021). Combined solar power and storage as cost-competitive and grid-compatible supply for China's future carbon-neutral electricity system. *Proceedings of the National Academy of Sciences*, 118(42), e2103471118.

[13] Cheng, Z., & Chow, M. Y. (2021). Resilient collaborative distributed AC optimal power flow against false data injection attacks: A theoretical framework. *IEEE Transactions on Smart Grid*, 13(1), 795-806.

[14] Sun, X., He, Y., Wu, D., & Huang, J. Z. (2023). Survey of distributed computing frameworks for supporting big data analysis. *Big Data Mining and Analytics*, 6(2), 154-169.

[15] Perera, C. (2024). Optimizing Performance in Parallel and Distributed Computing Systems for Large-Scale Applications. *Journal of Advanced Computing Systems*, 4(9), 35-44.

[16] Zhang, L., Peng, J., Zheng, J., & Xiao, M. (2023). Intelligent cloud-edge collaborations assisted energy-efficient power control in heterogeneous networks. *IEEE Transactions on Wireless Communications*, 22(11), 7743-7755.

[17] Deshmukh, S., Thirupathi Rao, K., & Shabaz, M. (2021). Collaborative learning based

straggler prevention in large-scale distributed computing framework. *Security and communication networks*, 2021(1), 8340925.

[18] Li, J., Gu, C., Xiang, Y., & Li, F. (2022). Edge-cloud computing systems for smart grid: state-of-the-art, architecture, and applications. *Journal of Modern Power Systems and Clean Energy*, 10(4), 805-817.

[19] Brynjolfsson, E., Jin, W., & McElheran, K. (2021). The power of prediction: predictive analytics, workplace complements, and business performance. *Business Economics*, 56(4), 217-239.

[20] Liu, Z., Huang, B., Li, Y., Sun, Q., Pedersen, T. B., & Gao, D. W. (2024). Pricing game and blockchain for electricity data trading in low-carbon smart energy systems. *IEEE Transactions on Industrial Informatics*, 20(4), 6446-6456.

[21] Hossain, M. A., Hossain, A. R., & Ansari, N. (2022). AI in 6G: Energy-efficient distributed machine learning for multilayer heterogeneous networks. *IEEE Network*, 36(6), 84-91.

[22] Du, X., Tang, S., Lu, Z., Gai, K., Wu, J., & Hung, P. C. (2022). Scientific workflows in iot environments: A data placement strategy based on heterogeneous edge-cloud computing. *ACM Transactions on Management Information Systems (TMIS)*, 13(4), 1-26.

[23] Sun, W., Li, Z., Wang, Q., & Zhang, Y. (2022). FedTAR: Task and resource-aware federated learning for wireless computing power networks. *IEEE internet of things journal*, 10(5), 4257-4270.

[24] Joshi, A., Capezza, S., Alhaji, A., & Chow, M. Y. (2023). Survey on AI and machine learning techniques for microgrid energy management systems. *IEEE/CAA Journal of Automatica Sinica*, 10(7), 1513-1529.