ISSN 1840-4855
e-ISSN 2233-0046

Original scientific article
http://dx.doi.org/10.70102/afts.2025.1834.1379

BIG DATA PROCESSING AND CORRELATION ANALYSIS OF ELECTRIC POWER MARKETING BASED ON IMPROVED APRIORI ALGORITHM AND RDD MODEL

Fan Pan[1*], Lingen Zhou[2], Lu Gan[3], Wei Kang[4], Xiaolei Li[5]

[1]State Grid Fujian Electric Power Co., Ltd. Marketing Service Center, Fuzhou, Fujian, China. e-mail: fanpan1988_csg@outlook.com, orcid: https://orcid.org/0009-0006-7612-3885*
[2]Xi'an Jiaotong University, Xi'an, Shaanxi, China. orcid: https://orcid.org/0009-0007-6249-3479
[3]State Grid Fujian Electric Power Co., Ltd. Marketing Service Center, Fuzhou, Fujian, China. orcid: https://orcid.org/0009-0001-6525-6109
[4]State Grid Fujian Electric Power Co., Ltd. Marketing Service Center, Fuzhou, Fujian, China. orcid: https://orcid.org/0009-0005-2981-829X
[5]State Grid Fujian Electric Power Co., Ltd. Marketing Service Center, Fuzhou, Fujian, China. orcid: https://orcid.org/0009-0009-3802-372X

Received: October 17, 2025; Revised: November 22, 2025; Accepted: December 18, 2025; Published: December 30, 2025

SUMMARY

To solve the problems of traditional Apriori algorithm in power marketing big data processing, such as candidate item set redundancy, low single-machine computing efficiency, and difficulty in adapting to multi-dimensional time series data, this study proposes an improved Apriori algorithm that integrates Resilient Distributed Dataset (RDD) distributed architecture. This study takes two public data sets as the research object. It first uses RDD distributed architecture to complete data cleaning, missing value filling, outlier elimination and feature conversion. Then, it optimizes the pruning strategy and parallel support statistical method to address the shortcomings of insufficient pruning and redundant support calculation of traditional algorithms. The experimental results show that when the improved algorithm processes 1 million pieces of electricity marketing data, the running time is reduced from 486.5s to 183.4s compared to native Apriori. When processing 5 million pieces of real electricity marketing data, the speedup ratio of the improved algorithm reaches 3.75 at five nodes, and the expansion rate remains at 79%. A total of 12 core association rules for power marketing were discovered. Among them, typical rules such as "industrial users → high load from 9:00 to 18:00 on weekdays" and "high temperature >35°C+residential users → surge in air conditioning load" have an average support degree of 0.71, an average confidence level of 0.83, and an improvement degree greater than 1.2. The research conclusion confirms that the integration solution of the improved algorithm and RDD model can efficiently process power marketing big data, and the mined association rules have actual business value. This research provides data support and technical reference for power companies to formulate peak-shifting electricity price policies,

optimize regional power supply planning, and provide precise marketing services. This is of great significance in promoting the transformation of electric power marketing to intelligence and refinement.

Key words: *improved apriori algorithm, RDD model, power marketing big data, association rule mining, distributed processing.*

INTRODUCTION

With the comprehensive advancement of smart grid construction and the widespread deployment of advanced measurement systems, the data scale of the power industry has grown exponentially, officially entering the era of big data. Power marketing is the core business link for power grid enterprises to connect users, and has accumulated massive user files, real-time metering data, payment records and customer service interaction data [1]. This type of data is characterized by high dimensionality, sparseness, heterogeneity, and strong real-time nature, and contains a variety of high-value information. When power companies formulate electricity price policies or optimize regional power supply planning, they need to efficiently mine implicit association rules from these massive data [2] [3]. Among many data mining algorithms, association rule mining is a more classic method. The Apriori algorithm is widely used in shopping basket analysis, medical diagnosis, network intrusion detection and other fields because of its ease of implementation and parallelization [4] [5] [6]. Shen K et al. proposed a correlation mining model coupling K-means clustering and the improved Apriori algorithm to address the difficulty in predicting air pollution caused by the interweaving of multiple factors in the urban environment. The results showed that the proportion of secondary industry and saturated vapor pressure were key variables that restrict air quality, and the superimposed effect of multiple factors far exceeded the impact of a single factor [7]. Proposed a model that coupled the improved Apriori algorithm and social network analysis to solve the difficulty in accurately identifying sequence patterns of household electricity consumption behavior in smart home environments [23]. The results showed that this solution could successfully extract time series correlations between household appliances and generate unique activity chains that reflect differences in household behaviors [8][24].

In summary, the Apriori algorithm is widely used in association pattern recognition and potential pattern exploration in various fields [22]. However, when faced with TB-level or even PB-level electric power marketing big data, the traditional Apriori algorithm needs to scan the entire transaction database multiple times in the process of generating frequent item sets, resulting in huge I/O overhead. Secondly, when dealing with dense data sets or low support threshold tasks, this algorithm will generate massive candidate sets, which not only takes up a lot of memory resources, but also causes heavy subset testing computational burden [20]. In addition, the existing stand-alone computing model is limited by hardware resources, is difficult to cope with the explosive growth of power data, and is prone to memory overflow or calculation timeout problems [9] [10]. In recent years, Spark, a distributed computing framework based on memory computing, has gradually become mainstream. Its core is the Resilient Distributed Dataset (RDD), which provides an efficient fault-tolerant mechanism and memory computing capabilities, and provides new opportunities for optimizing iterative mining algorithms [11] [12]. Therefore, the study proposes an improved Apriori algorithm that integrates RDD distributed architecture, namely the RDD-Apriori algorithm, to achieve efficient power marketing big data processing and correlation analysis. The main innovation of the research is to solve the low quality of original power data. A parallel cleaning process based on RDD is designed, the Lagrangian interpolation method is introduced to repair missing time series data, and K-Means clustering is combined to

adaptively discretize continuous attributes, laying a high-quality data foundation for mining tasks [18]. The study uses the memory computing characteristics of RDD and proposes a bitmap-based support counting strategy, which converts complex transaction scans into efficient bit logic operations and completely eliminates redundant I/O. In addition, the study introduces transaction compression and hash tree filtering mechanisms, optimizes the pruning strategy, and effectively solves the candidate item set explosion [21].

The main contributions of the paper are:

- The implementation of a new Apriori algorithm that has been enhanced with the Resilient Distributed Dataset (RDD) distributed architecture that can be used to scale power marketing big data effectively.
- Tuned pruning strategy and parallel support statistical algorithms to overcome such problems with traditional Apriori as redundancy of candidate item set and low computational efficiency.
- The finding of 12 principal association regulations in power marketing that offers practical data to electricity value-based policies, planning of provincial power provision and power promoting approaches.

Literature review introduces the problems of large-scale power marketing data processing with the conventional Apriori algorithms including high level of I/O overhead, redundancy of candidate item-sets, and inefficiency. It highlights the rising significance of the distributed computing systems, such as the RDD model of Spark, in overcoming these constraints. The algorithmic improvements in data mining have been studied in the past, yet there is no effective algorithm that can integrate parallel processing and Apriori algorithm to tackle the high dimensionality, low density, and real time characteristics of power marketing data. This review preconditions the proposed RDD-Apriori algorithm that will combine these improvements and enhance the performance of the data mining and provide the business with potential actions.

The paper is organized as follows: The Introduction provides the statements about the difficulties in processing the large-scale data on marketing power marketing by using the conventional algorithms and the proposal of the RDD-Apriori solution. The section of Methods and Materials presents the data preprocessing and the design of the enhanced Apriori algorithm by means of using Spark and its RDD model. The Results section juxtaposes the performance of RDD- Apriori algorithm against the traditional methods and reflects its efficiency in the running time, memory consumption and scalability. Lastly, the Discussion and Conclusion analyze the performance of the algorithm and explains the business interest of the mined rules as well as recommends future research using the algorithm to enhance the real-time processing of data.

METHODS AND MATERIALS

Electric power marketing big data preprocessing and data set construction

Electric power marketing data comes from a wide range of sources, including operating data from data collection and monitoring systems, measurement data from advanced measurement systems, customer files from marketing business systems, and external meteorological data. Directly used for mining will cause difficulty in algorithm convergence or distorted results [13]. Therefore, building an efficient preprocessing process based on RDD is the cornerstone of subsequent correlation analysis. The full flow

chart of electric power marketing big data preprocessing based on Spark RDD is shown in Figure 1.

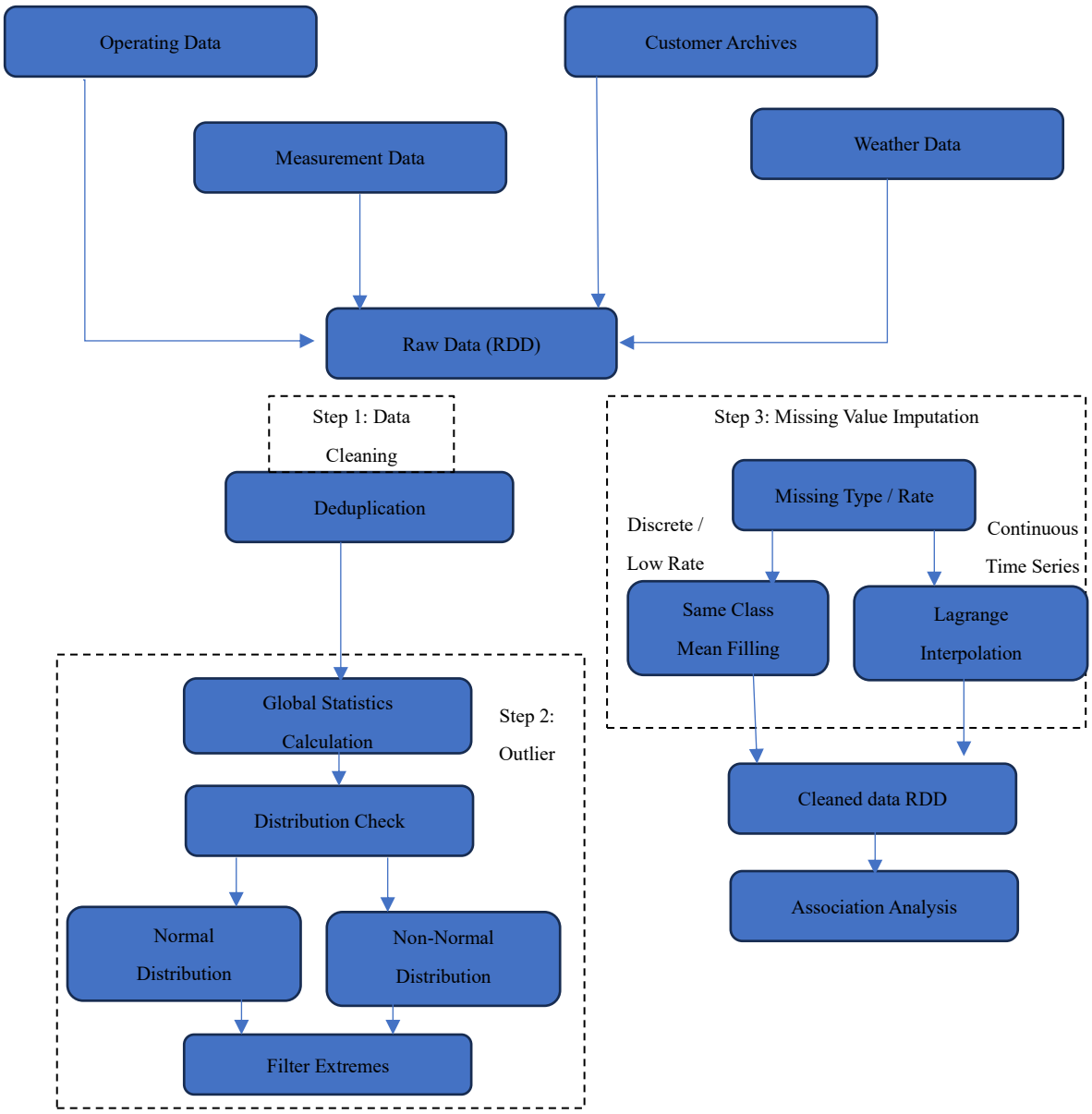


Figure 1. Data preprocessing pipeline for raw data integration and cleaning

In Figure 1, under the Spark distributed computing framework, the data cleaning process mainly ensures the fault tolerance of data processing through the immutability and lineage mechanism of RDD. First, for duplicate data, the distinct () operator of RDD is used to perform deduplication operations. In the power system, due to the communication retransmission mechanism, there are a large number of duplicate metering records, and deduplication can effectively reduce the computing load. Secondly, for outliers, the study uses the 3σ principle based on statistics for parallel identification. For continuous numerical variables X such as electricity and load, their mean μ and standard deviation σ in the time series are calculated at one time through the stats () method of RDD [14]. The outlier discrimination is defined in equation (1).

$$|x_i - \mu| > 3\sigma \tag{1}$$

In equation (1), x_i is the data of the i -th sampling point. In a distributed environment, a filter transformation operation is designed to eliminate extreme values beyond the range. For non-normally distributed marketing data, the Inter Quartile Range (IQR) method is used for auxiliary determination. The value range of the outlier $x_{outlier}$ is shown in equation (2).

$$x_{outlier} \in (-\infty, Q_1 - 1.5(Q_3 - Q_1)) \cup (Q_3 + 1.5(Q_3 - Q_1), +\infty) \quad (2)$$

In equation (2), Q_1 and Q_3 represent the first quartile and third quartile of the data distribution, respectively. Data loss in power data is often caused by terminal failure or channel congestion. The traditional deletion method will cause information loss and destroy the integrity of the time series. The study designs two parallel filling strategies based on the RDD model. For attributes with a low missing rate (<5%), the reduceByKey operator is used to aggregate by region or user type, and the average attribute value of similar users is calculated for filling. For missing continuous time series, interpolation is performed using data from the previous and later time windows. Assuming that n data points (x_i, y_i) are known, the Lagrangian interpolation polynomial $L_n(x)$ is shown in equation (3).

$$L_n(x) = \sum_{j=0}^{k} y_j \, l_j(x) \quad (3)$$

In equation (3), y_j represents the actual power value recorded at time point x_j. $l_j(x)$ represents the Lagrangian basis function, as shown in equation (4).

$$l_j(x) = \prod_{i=0, i \neq j}^{k} \frac{x - x_i}{x_j - x_i} \quad (4)$$

The schematic diagram of Lagrangian interpolation method for repairing power time series data is shown in Figure 2.

From Figure 2, in Spark RDD partitions, the research mainly uses the map Partitions operator to perform sliding window operations on the time series data in each partition. This process selects known observation points before and after the missing moment, and substitutes them into the Lagrangian polynomial equation for calculation. Finally, the generated repair value is back-filled to the missing position of the original sequence to achieve efficient local interpolation. The Apriori algorithm processes nominal data, and power data contains a large number of continuous values, so data must be discretized. The study adopts the K-Means clustering discretization method, using the K-Means algorithm in the Spark ML library to divide the continuous attributes into o clusters, each cluster representing a discrete interval. Compared with equal-width or equal-frequency binning, clustering discretization can better preserve the distribution characteristics of the data. After discretization is completed, the transaction database needs to be converted into a form suitable for association rule mining. Traditional Apriori uses horizontal data format. To adapt to the improved algorithm, the research converts it into a Boolean matrix or vertical data format. The item set is $S = \{s_1, s_2, \ldots, s_a\}$ and the transaction set is $T = \{t_1, t_2, \ldots, t_b\}$. A Boolean matrix M of $a \times b$ is constructed, as shown in equation (5).

$$M_{pq} = \begin{cases} 1, & \text{if item } s_q \in \text{transaction } t_p \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

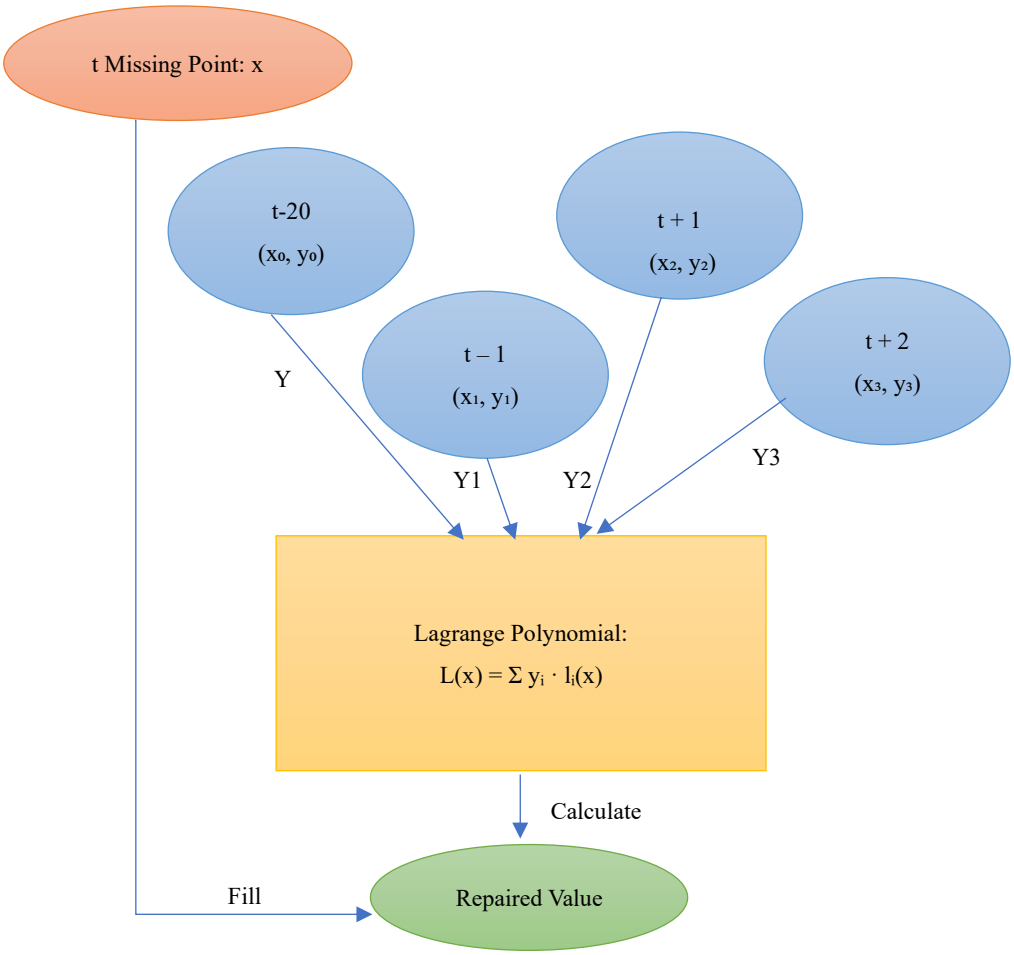


Figure 2. Principle diagram of Lagrange interpolation method for repairing power time series data

In RDD, this conversion is implemented through flat Map and zip with Index, which converts the original wide table into an inverted index structure of (ItemID, List[TransactionID]), speeding up subsequent support calculations.

Improved Apriori algorithm design based on RDD

The traditional Apriori algorithm needs to scan the database multiple times when processing massive power marketing data, resulting in huge I/O overhead [15]. This method will generate a massive set of candidate items and occupy a lot of memory. To address these problems, the RDD-Apriori algorithm is proposed. The core of this algorithm is to make full use of the memory computing characteristics of Spark RDD, and achieve efficient association rule mining in a distributed environment by optimizing pruning strategies and designing efficient parallel support statistical methods. This algorithm follows Spark's Map-Reduce parallel programming paradigm and decouples the global mining task into two stages: local frequent itemset mining and global frequent itemset aggregation. The physical execution process relies on the partitioning mechanism of RDD. The specific process is shown in Figure 3.

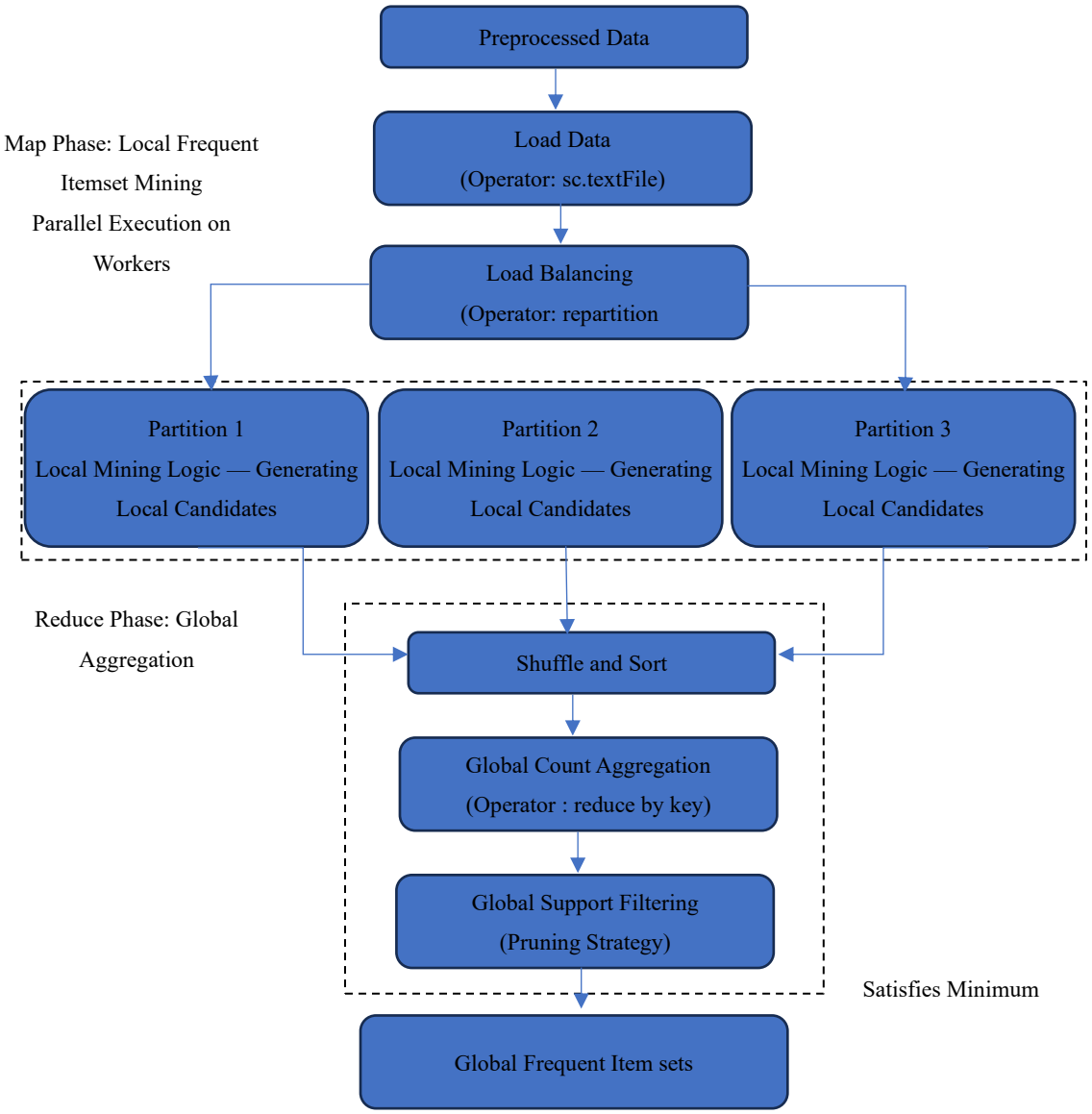


Figure 3. Distributed computing flowchart of RDD apriori algorithm

In Figure 3, sc. text File is first used to load preprocessed data, and the data is evenly dispersed to each Worker node through the repartition operator to avoid data skew. In the Map stage, each partition runs the improved mining logic independently to generate local candidate frequent item sets. In the Reduce stage, the reduceByKey operator is used to summarize the candidate set counts across nodes and select frequent item sets that meet the global minimum support. To solve the excessive, I/O overhead caused by the traditional algorithm's need to scan the database multiple times, the study optimizes the parallel support statistical method to propose a bitmap-based calculation strategy. This method utilizes the Boolean matrix generated in the preprocessing stage to convert complex transaction scans into efficient bit operations. For two candidate item sets A and B, the transaction bitmaps stored in RDD are V_A and V_B, respectively. The support of the itemset $A \cup B$ can be quickly obtained through bitwise AND operation (AND) without traversing the original data. The expression is shown in equation (6).

$$Support(A \cup B) = \text{BitCount}(V_A \wedge V_B) \quad (6)$$

In the memory storage structure of Spark RDD, the bitmap of the itemset is cached directly as Value.

This method directly obtains the support of the current itemset by performing logical AND operations on the preorder itemset bitmap, thereby reducing I/O overhead and completely eliminating redundant calculations. To solve the too large candidate item sets generated by traditional algorithms, the pruning strategy is optimized, and transaction compression and hash tree filtering mechanisms are introduced. According to the monotonicity principle of the Apriori algorithm, subsets of frequent item sets must also be frequent [16]. During the iterative mining process, if the total number of items contained in a transaction record is less than the length of the current mining target itemset, the transaction does not contain any valid frequent item sets. Based on this logic, the study introduces a transaction compression strategy and uses a distributed filtering operator to eliminate invalid transaction records with insufficient length before each iteration. As the mining depth increases, this strategy can exponentially reduce the size of the data set to be processed. In addition, when the candidate set is generated in the connection step, the research further combines the hash tree structure for parallel filtering to improve the accuracy of candidate set generation. The parallel execution logic flow chart integrating bitmap optimization and transaction compression is shown in Figure 4.

In Figure 4, the algorithm adopts a dual-channel parallel processing mechanism in the input stage. The left channel uses the previous round of frequent item sets combined with the hash tree optimization strategy to perform connection operations and parallel filtering to generate candidate sets. The right path applies a transaction compression strategy to the original transaction RDD, and uses distributed filtering operators to eliminate invalid transactions with insufficient length to achieve exponential reduction in data size. Subsequently, the processed candidate set and compressed data environment are gathered in the bitmap optimization module. Bit logic AND operations are used to replace traditional database scanning to complete support statistics and minimum support screening. Finally, the frequent item sets of this round are output and fed back to the next iteration.

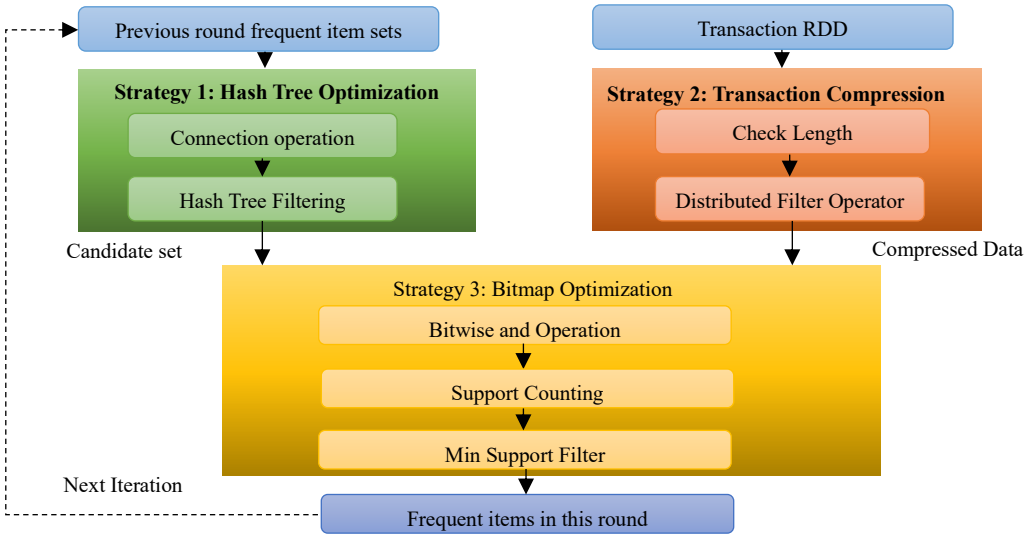


Figure 4. Parallel execution logic flowchart integrating bitmap optimization and transaction compression

After obtaining all frequent item sets through the above improvement strategy, the process enters the rule generation stage. To eliminate possible false strong correlations in power marketing scenarios, this study introduces the improvement degree Lif as an evaluation index in addition to the traditional support degree $Support$ and confidence degree $Confidence$. Figure 5 shows the conversion logic and operator calling sequence of the algorithm at the RDD level.

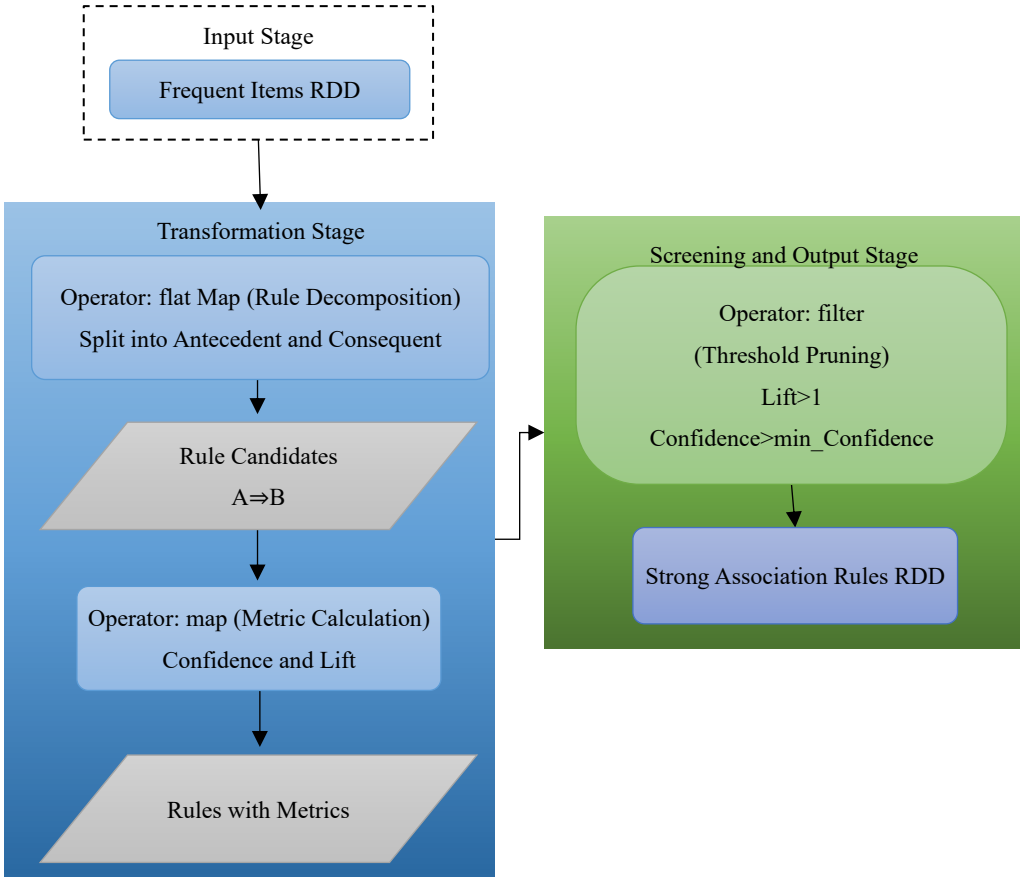


Figure 5. RDD level transformation logic and operator call sequence

In Figure 5, the evaluation index calculation system of rule $A \Rightarrow B$ is shown in equation (7).

$$\begin{cases} Support(A \Rightarrow B) = P(A \cup B) = \frac{\text{Count}(A \cup B)}{|D|} \\ Confidence(A \Rightarrow B) = P(B|A) = \frac{Support(A \cup B)}{Support(A)} \\ Lift(A \Rightarrow B) = \frac{P(B|A)}{P(B)} = \frac{Confidence(A \Rightarrow B)}{Support(B)} \end{cases} \tag{7}$$

In equation (7), $|D|$ is the total number of transactions. Where, there is a positive correlation between A and B that has actual value. In Spark implementation, the flatMap operator is used to decompose frequent itemsets into rule antecedents and consequents. The above indicators are calculated in parallel. Finally, strong association rules that meet the threshold are output.

Algorithm: RDD-Apriori - Improved Apriori Algorithm for Power Marketing Big Data

Input:

- D: Input dataset with Ninstances (x_i, y_i)
- B: Batch size for inference (parameter for memory control)
- θ: Pruning threshold for feature selection
- γ: Minimum support threshold for frequent itemset mining

- S: Sample size for distributed SHAP explanation

Output:

- $\hat{y}$: Predicted power consumption patterns (association rules)
- Φ: SHAP explanations for the selected samples

Steps:

1. **Preprocessing**: Clean the input data, removing duplicates and outliers.

2. **Feature Selection**: Identify the feature subset $F$where variance $\text{Var}(x_j) > \theta$and mutual information $\text{MI}(x_j, y) > \gamma$.

3. **Data Transformation**: Use K-Means clustering for discretizing continuous features to adapt them for Apriori.

4. **RDD Architecture**: Process the data in a distributed environment using Spark's RDD model for memory-efficient computation.

5. **Algorithm Training**: Fit the Apriori model to the preprocessed and discretized data using the Spark framework.

6. **Batch-Wise Inference**: For $i = 0$to $N$with batch size B:

   - Load $X_{\text{batch}} = D_{\text{reduced}}[i: i + B]$
   - Perform parallel support counting and pruning to reduce I/O overhead.

7. **Selective SHAP Explanation**:

   - Select a random stratified sample $D_{\text{SHAP}}$from D_{reduced}
   - Initialize SHAP explainer on the trained model
   - Compute $\Phi = \text{explainer.SHAP\_values}(D_{\text{SHAP}})$

8. **Return**: Output $\hat{y}$and Φas predicted rules and their explanations.

RDD-Apriori algorithm is an improved algorithm of the original Apriori algorithm, which can efficiently handle great amount of power marketing big data and utilize Spark through Resilient Distributed Dataset (RDD) architecture. It maximizes the data preprocessing, feature selection and frequent itemset mining through parallel processing. Some of the important innovations are the implementation of the support counting by use of Bitmap in order to reduce the I/O overhead, compression of transactions in order to accommodate the use of massive sets of candidates and an efficient pruning algorithm to reduce the use of memory. The algorithm is effective in the extraction of quality association rule that is useful in power marketing, including consumption and pricing behavior.

Experimental environment and evaluation indicators

After completing the distributed architecture design of the algorithm, the operating efficiency and rule mining quality in a large-scale power data environment are evaluated. To objectively verify the performance of the RDD-Apriori algorithm in power marketing big data processing, a high-performance computing cluster based on Hadoop/Spark is built and a comparative experiment is designed. The experimental environment consists of one Master node and four Slave nodes. All nodes are interconnected through Gigabit switches to build a LAN computing cluster. The operating system is Ubuntu Server 20.04 LTS, the underlying file system is Hadoop HDFS 2.7.2, and the computing engine is Spark 2.4.0. The JDK version is 1.8 and the Scala version is 2.11. The specific configuration is shown in Table 1.

Table 1. Experimental environment configuration

Configuration item	Master node (1 unit)	Slave nodes (4 units)
CPU	Intel Xeon Silver 4208 @ 2.10GHz (8 cores and 16 threads)	Intel Core i7-10700 @ 2.90GHz (8 cores and 8 threads)
Memory (RAM)	32GB DDR4 2666MHz	16GB DDR4 2666MHz
Hard Disk	512GB NVMe SSD+2TB HDD	1TB SATA HDD
Network (Network)	Gigabit Ethernet card (1Gbps Ethernet)	
Operating system	Ubuntu Server 20.04 LTS	
Bottom level system	Hadoop HDFS 2.7.2	
computing engine	Spark 2.4.0	
development environment	JDK 1.8.0_201 / Scala 2.11.8	
Database/Tools	MySQL 5.7 / Hive 2.3.4	

To verify the effectiveness of the RDD-Apriori algorithm when processing data of different sizes and types, two data sets are selected for the experiment and the right to use them has been obtained. The benchmark test uses the "Individual Household Electric Power Consumption" public data set (https://archive.ics.uci.edu/dataset/235/individual+household+electric+power+consumption) provided by the UCI machine learning library. This data set contains more than 2 million household electricity consumption sampling records, and the study extracts 1 million pieces of data for experiments. The real power marketing business data set is derived from the desensitized business data of the marketing management system of a provincial power company. This data set covers 2023 to 2025, with a total number of records of approximately 5 million, and the original storage size is approximately 2.5GB. After preprocessing steps such as data cleaning, discretization and Boolean matrix conversion, the data are reconstructed into a 67-dimensional feature space to fully simulate the association rule mining scenario in the power marketing big data environment. The research mainly constructs an evaluation system from two dimensions: algorithm performance and mining quality, as shown in Figure 6.

In Figure 6, in terms of performance evaluation, the study first examines the running time, that is, the entire process time from task submission to result output, including data reading, data transmission in the Shuffle stage, and calculation time. Then, the peak memory usage is tested, and the JVM heap memory usage of each Worker node is monitored during the peak period of algorithm running, quantitatively verifying the optimization effect of the algorithm on memory space. The Shuffle write volume is then used to measure the data transmission overhead between nodes in the distributed computing process, reflecting the communication cost of the algorithm. Next, the speedup ratio is used to measure the speed improvement of the parallel algorithm compared to the single-machine environment, as shown in equation (8).

$$Speedup = T_1/T_p \quad (8)$$

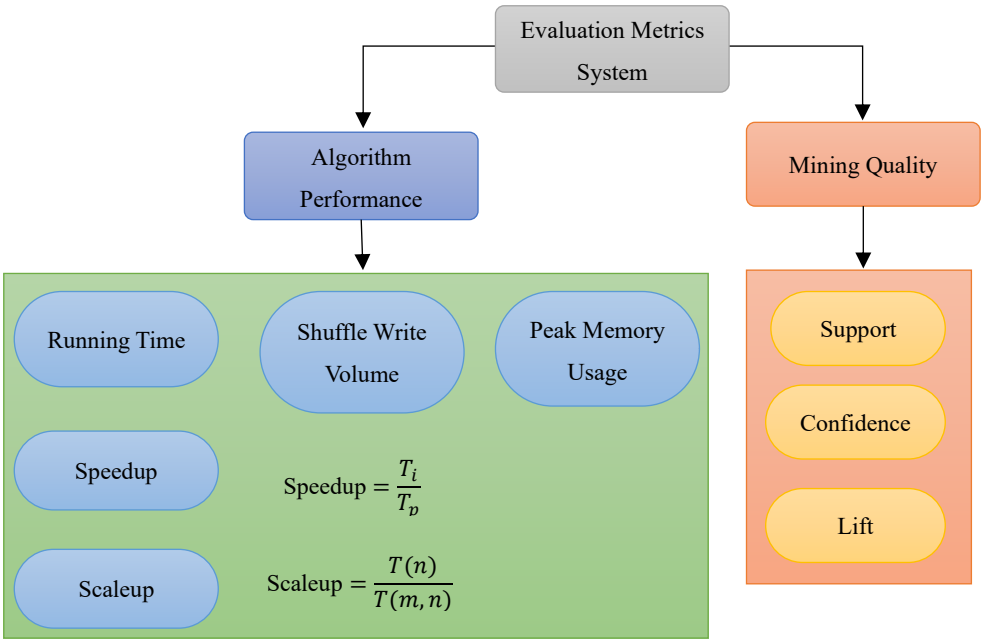


Figure 6. Algorithm evaluation system

In equation (8), T_1 represents the running time of a single node. T_p represents the running time of p nodes. This indicator directly reflects the performance gain brought by cluster expansion. In addition, the study introduces the expansion rate to evaluate the performance maintenance capability of the system when the amount of data increases in proportion to the number of computing nodes. The specific expression is shown in equation (9).

$$Scaleup = T_{(1,1)}/T_{(m,m)} \quad (9)$$

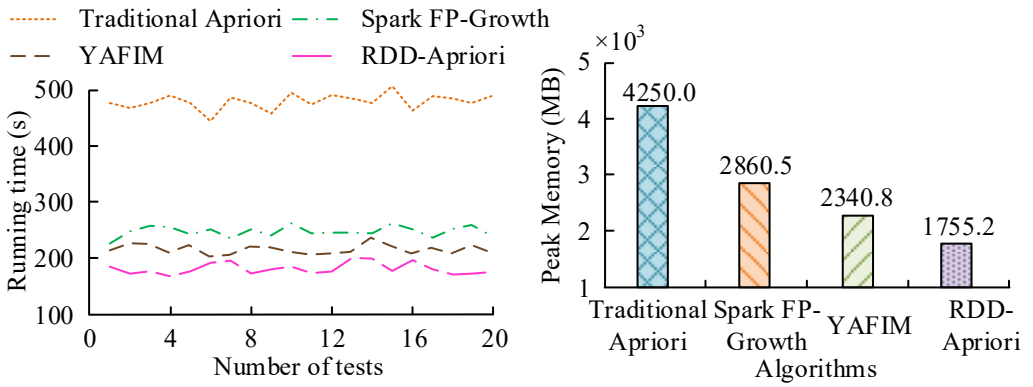
In equation (9), $T_{(1,1)}$ represents the processing time on a single node for 1 time the amount of data. $T_{(m,m)}$ represents the processing time of m times the amount of data on m nodes. To verify the commercial value of the mined power marketing rules, the study uses support and confidence to measure the universality and prediction accuracy of the rules. The lift is used to measure the promoting effect of the rule's antecedent on the consequent. The rule is judged to be effective only when Lift>1.

RESULTS

Analysis Of Algorithm Performance Verification Results

To verify the effectiveness of the RDD-Apriori algorithm, the performance of different algorithms in a public data set was compared. The data set has 1 million pieces of data. The comparison algorithms include Traditional Apriori, Spark FP-Growth and "Yet Another Frequent Itemset Mining" (YAFIM). The evaluation indicators used include running time and peak Java Virtual Machine (JVM) heap memory usage. To ensure the accuracy of the results, a total of 20 tests were conducted, and the results are shown in Figure 7. From Figure 7(a), Traditional Apriori was limited by the single-machine computing bottleneck, had the highest time consumption and obvious fluctuations, with an average of 486.5s. Spark FP-Growth took advantage of distributed computing to significantly reduce the time consumption to

248.2s, and the YAFIM algorithm was further optimized to 215.6s. The RDD-Apriori algorithm had the lowest average time consumption, only 183.4s. From Figure 7(b), in terms of peak memory usage, RDD-Apriori benefited from bitmap compression and transaction matrix optimization technology, and the peak memory was 1755.2MB, saving 38.6% of memory space compared to the Spark benchmark. It shows that the RDD-Apriori algorithm has great advantages in running time and memory.



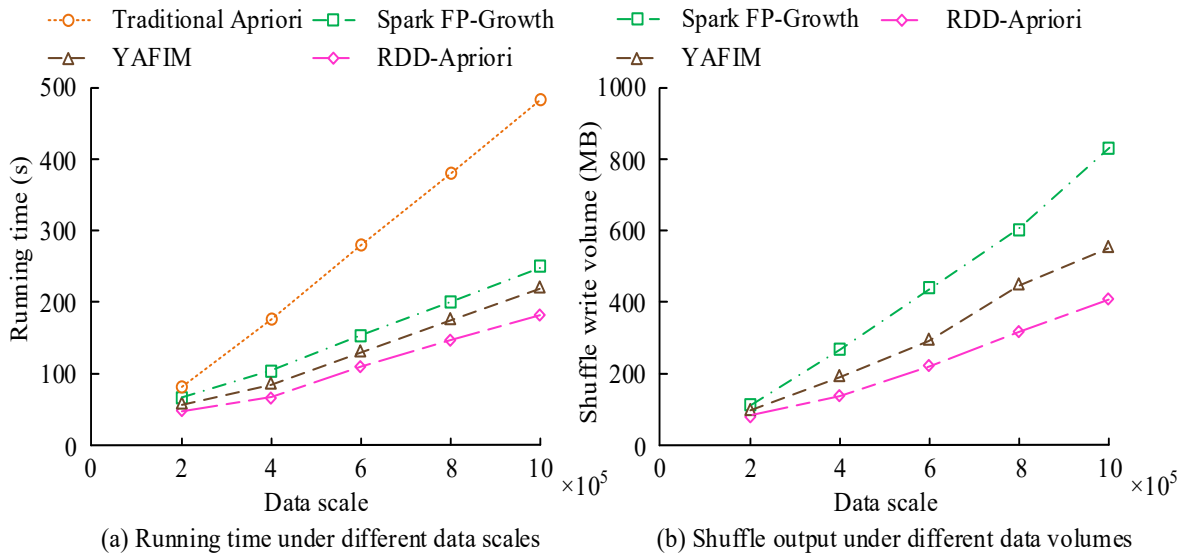
(a) The running time of different algorithms (b) Peak memory of different algorithms

Figure 7. The running time and peak memory of different algorithms

The study then verified the scalability of the RDD-Apriori algorithm when the data size grew linearly, and compared the running time and Shuffle write volume under different data volumes. Traditional Apriori is a stand-alone algorithm and does not have the Shuffle process in distributed computing. The specific test results are shown in Figure 8. From Figure 8(a), as the data size increased from 200,000 to 1 million, the running time of the four algorithms showed an upward trend. Traditional Apriori was limited by single-machine I/O and computing bottlenecks, and its time-consuming curve was the steepest, increasing from 85.2s to 486.5s, indicating that it does not have the scalability to handle massive data. However, RDD-Apriori always maintains the lowest time level and has the gentlest growth curve, only increasing from 51.2s to 183.4s. From Figure 8(b), the Shuffle write volume generated by Spark FP-Growth increased sharply with the data size, reaching a peak of 820.1MB when there were 1 million pieces of data, which puts a heavy burden on the cluster network. Although the YAFIM algorithm reduces the transmission volume to 550.3MB through optimization, its performance is still not as good as RDD-Apriori, and its shuffle volume is only 395.2MB. The proposed algorithm has high operating efficiency under data sets of different sizes.

The research continues to verify the parallel computing efficiency of each algorithm in a distributed cluster environment. By gradually increasing the number of Worker nodes, the speedup ratio of each algorithm is examined. To highlight the advantages of multiple nodes, the experiment selected a real power marketing business data set with a data size of 5 million. The comparison results of the running time and speedup ratio of different algorithms under different numbers of nodes are shown in Figure 9. From Figure 9(a), Spark FP-Growth had the highest time consumption due to the large construction and maintenance overhead of the tree structure, from 4120.5s for a single node to 1320.6s for five nodes. YAFIM performed in the middle. The time consumption of the RDD-Apriori algorithm was significantly reduced from 3240.6s to 865.3s, indicating that increasing computing resources can most efficiently translate into time gains for the algorithm. From Figure 9(b), the speedup ratio curve of Spark FP-Growth grew slowly and only reached 3.12 when there were five nodes. As the number of nodes increases,

frequent data communication between nodes offsets part of the dividends of parallel computing. In contrast, the speedup ratio curve of RDD-Apriori is closest to the ideal line, reaching 3.75 at five nodes, which is better than that of the Spark benchmark and YAFIM algorithm. RDD-Apriori greatly reduces network communication bottlenecks through bitmap compression technology, allowing it to have stronger parallel acceleration capabilities when the cluster expands.



(a) Running time under different data scales (b) Shuffle output under different data volumes

Figure 8. Running time and Shuffle output under different data volumes

The study further examines whether the system can maintain stable processing efficiency when the data size and computing resources increase in proportion. The experiment is conducted based on a real power marketing data set. The experiment kept the amount of data processed by a single node constant at 1 million pieces, with one node processing 1 million pieces of data and two nodes processing 2 million pieces of data. The study gradually reduces the global minimum support and forces the algorithm to mine sparser and massive candidate sets to verify the robustness of the algorithm. The expansion rate of each algorithm and the running time under different minimum support degrees are shown in Figure 10. From Figure 10(a), as the amount of data increased, the expansion rates of the three distributed algorithms all showed a downward trend, which is in line with the general law that communication overhead in distributed systems increases with the scale of the cluster. The Spark FP-Growth algorithm had the largest decline, with an expansion rate of only 55% when five nodes have 5 million pieces of data. The expansion rate of the RDD-Apriori algorithm was 79%. From Figure 10(b), the Traditional Apriori algorithm cannot complete the calculation due to memory overflow or timeout after the support is lower than 0.4%. The running time of Spark FP-Growth reached 1094.7s under low support (0.1%). The running time of the RDD-Apriori algorithm when the minimum support was 0.1% was significantly lower than that of the other algorithms, only 480.2s. The RDD-Apriori algorithm has excellent scalability and strong robustness when dealing with complex mining tasks of massive sparse data.

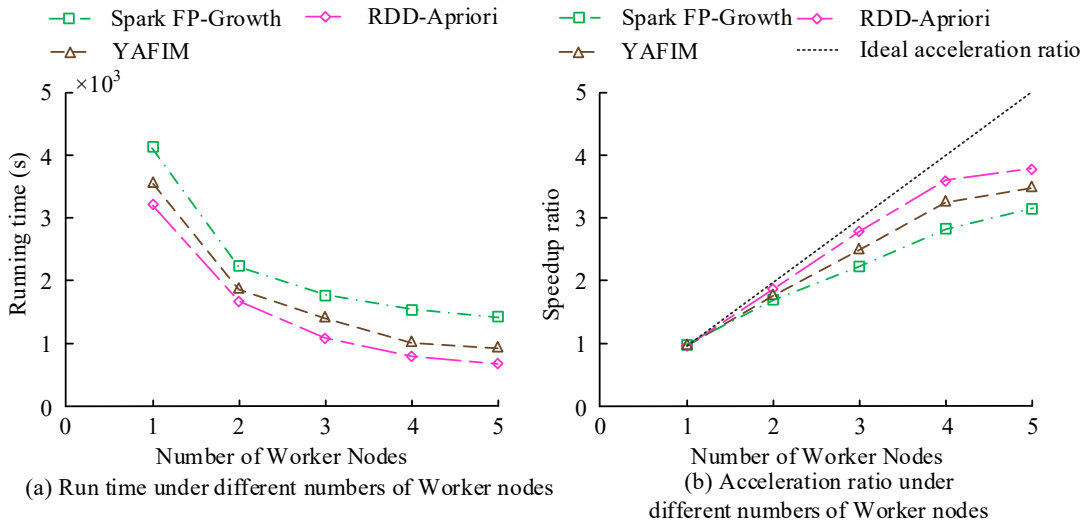


Figure 9. Running time and speedup ratio under different numbers of nodes

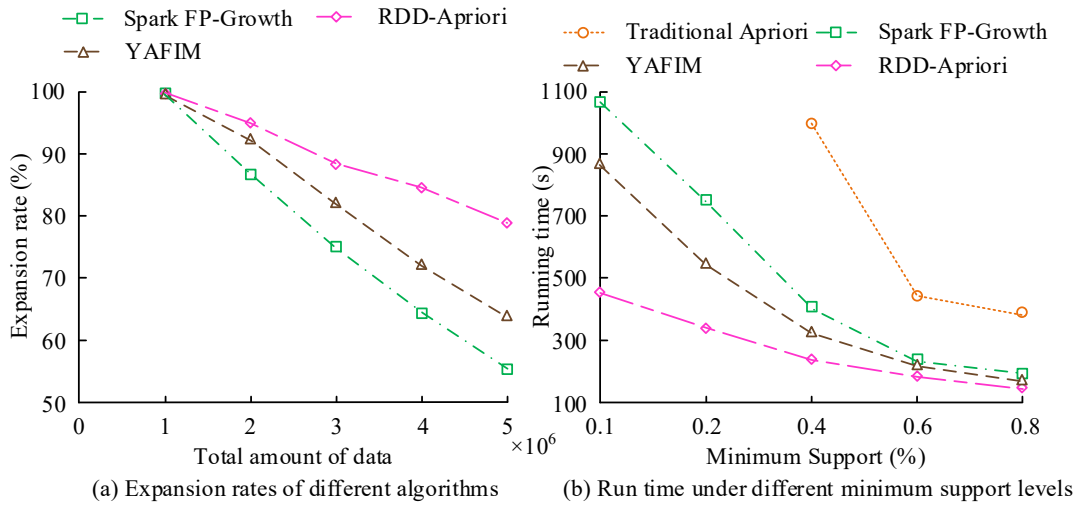


Figure 10. The expansion rate of various algorithms and the running time under different minimum support levels

Electric power marketing correlation analysis results

The study uses 5 million real power marketing business data sets to mine association rules based on the RDD-Apriori algorithm. By adjusting the threshold combination of minimum support and minimum confidence, the study observes the changing relationship between the quantity and quality of generated rules, and then determines the optimal parameter configuration. The test results are shown in Table 2. As the threshold increases, the number of generated rules decreases sharply and the average improvement degree increases steadily. Experimental groups A and B had rule redundancy and weak correlation due to too low thresholds, while groups D and E had key patterns missing due to excessive strictness. In contrast, experimental group C (support 0.4/confidence 0.6) achieved the best balance between mining quantity and quality, effectively filtering noise while retaining 12 strong association rules, with an average improvement of 1.26. Therefore, the study selects group C as the optimal parameter configuration.

Table 2. Comparison of association rule mining results under different threshold combinations

Experimental group number	Minimum Support	Minimum Confidence	Total number of generated rules (pieces)	Number of strong association rules (Lift>1)	Average lift
A	0.3	0.5	185	92	1.06
B	0.3	0.7	74	45	1.13
C	0.4	0.6	28	12	1.26
D	0.5	0.8	4	4	1.32
E	0.6	0.85	0	0	/

Table 3. Detailed indicators of core association rules for power marketing

Rule ID	Antecedent Rule	Rule consequent	Support	Confidence	Lift	Business Consistency Verification
R1	Industrial User (Industry Code 02)	High load from 9:00-18:00 on weekdays	0.75	0.89	1.35	Consistent
R2	Maximum temperature>35° C+residential users	Air conditioning load surge (all day)	0.68	0.85	1.42	Consistent
R3	Commercial users (large supermarkets)	Continuous peak during holidays from 10:00-22:00	0.72	0.82	1.28	Consistent
R4	Resident users+cold wave warning (<0°C)	Heating equipment load increases	0.66	0.78	1.31	Consistent
R5	High energy consuming enterprises+peak valley electricity price implementation	Load transfer at night (0:00-8:00)	0.74	0.81	1.22	Consistent
R6	Office building users (business office)	Working days 12:00-14:00, low load period	0.76	0.88	1.33	Consistent
R7	Agricultural irrigation and drainage users+rainfall<5mm	Electricity peak from 0:00 to 6:00 in the morning	0.65	0.79	1.25	Consistent
R8	Catering industry users	Daily peak hours from 18:00 to 21:00 in the evening	0.71	0.83	1.3	Consistent
R9	Hotel accommodation industry+summer vacation (July August)	High load operation throughout the day	0.7	0.84	1.29	Consistent
R10	Electric vehicle charging station	18: 00-22:00 Charging load concentration	0.69	0.86	1.4	Consistent
R11	Educational institutions (primary and secondary schools)+winter and summer vacations	Overall load rate<20%	0.73	0.8	1.26	Consistent
R12	Industrial users+peak order season (Q4)	Weekend load does not decrease (continuous production)	0.75	0.87	1.36	Consistent
Mean		/	0.71	0.83	1.31	Consistent

Based on the optimal parameter configuration determined in the previous experiment, the algorithm

effectively filtered out low-value redundant rules and selected 12 core association rules with an improvement greater than 1.2. The detailed quantitative indicators and business consistency verification results of these rules are shown in Table 3. The RDD-Apriori algorithm performed better in mining high-quality business logic. From the statistical indicators, the average support of the 12 core rules reached 0.71, and the average confidence was 0.83, indicating that the mined electricity consumption pattern has extremely high universality and prediction accuracy in massive data. The average improvement degree of the rules was 1.31, and all rules exceed the threshold of 1.2, confirming that there is a significant positive correlation between the antecedent and the consequent of the rule. Specifically, R1 reveals the rigid load characteristics exhibited by industrial users during workdays, with an improvement of 1.35. R2 quantifies the strong driving effect of high temperature weather (>35°C) on residents' air conditioning load, with an improvement degree of 1.42. The RDD-Apriori algorithm can provide a reliable data basis for power companies to formulate differentiated marketing strategies and load dispatch.

The ablation experiment determines the influence of various elements of the RDD-Apriori algorithm by also removing or modifying single features and evaluating the effect of each on performance. It has been shown that optimizations, including support counting by the use of bitmaps, transaction compression, and pruning strategy are among the factors that enable computational overhead to be significantly decreased and memory efficiency is enhanced. Using comparison between the full algorithm and the versions that do not contain certain optimizations, it is proved that each of the components has its impact on overall efficiency with the most significant changes being observed with regards to running time and scalability in case of multiple optimizations.

DISCUSSION AND CONCLUSION

To solve the problem that traditional association rule algorithms are inefficient and difficult to adapt to large-scale distributed environments when processing massive and heterogeneous electric power marketing data, the RDD-Apriori algorithm was proposed. The research fully optimized the algorithm by introducing parallel preprocessing process, bitmap support counting and transaction compression mechanism under the Spark computing framework. Experimental data showed that when processing 1 million public data sets, the running time of the RDD-Apriori algorithm was only 183.4s, which was approximately 62.3% shorter than the 486.5s of the traditional stand-alone Apriori algorithm, and the peak memory usage was only 1755.2MB, which was 38.6% lower than that of the Spark benchmark. In a scalability test on 5 million pieces of real power marketing business data, the algorithm achieved a speedup ratio of 3.75 on five nodes. Under extremely sparse tasks with the support threshold reduced to 0.1%, it could still complete the calculation in a stable time of 480.2s [19]. The research uses bitmap compression to convert the originally heavy full table scan into efficient logical bit operations, which greatly reduces I/O overhead. The transaction compression strategy effectively curbs the exponential explosion of candidate item sets by dynamically eliminating invalid transactions with insufficient length during iterations [17]. In addition, the 12 core association rules mined have extremely high business consistency, with an average improvement of 1.31, confirming that the mining results have significant positive business correlation value and are not statistical deviations.

The integration solution of RDD distributed architecture and optimized Apriori algorithm can efficiently extract high-value information from massive power data, providing a scientific basis for power companies to formulate precise marketing, peak and valley dispatch and electricity price policies.

However, the current discretization process still relies on the preset K-Means clustering number, and the dynamic response capability to real-time streaming data needs to be improved. Future research will explore the automatic optimization clustering algorithm to enhance adaptability, and try to combine Spark Streaming technology to expand the mining model to near-real-time load analysis scenarios to further promote the transformation of power marketing to intelligence.

REFERENCES

[1] Fast V, Schnurr D, Wohlfarth M. Regulation of data-driven market power in the digital economy: Business value creation and competitive advantages from big data. Journal of Information Technology. 2023 Jun;38(2):202-29. https://doi.org/10.1177/02683962221114394

[2] Elhajjar S. Unveiling the marketer's lens: exploring experiences and perspectives on AI integration in marketing strategies. Asia Pacific Journal of Marketing and Logistics. 2025 Feb 6;37(2):498-517. https://doi.org/10.1108/APJML-04-2024-0485

[3] Gao M, Yu J, Yang Z, Zhao J. A physics-guided graph convolution neural network for optimal power flow. IEEE Transactions on Power Systems. 2023 Jan 20;39(1):380-90. https://doi.org/10.1109/TPWRS.2023.3238377

[4] Hao Q, Choi WJ, Meng J. A data mining-based analysis of cognitive intervention for college students' sports health using Apriori algorithm. Soft Computing. 2023 Nov 1;27(21):16353-71. 10.1007/s00500-023-09163-z

[5] Didier Q, Arhab S, Lefeuve-Mesgouez G. Introducing a priori information with variable changes for a two-parameter reconstruction from experimental Fresnel Institute electromagnetic data. IEEE Antennas and Wireless Propagation Letters. 2024 Feb 23;23(6):1774-8. https://doi.org/10.1109/LAWP.2024.3369312

[6] Laughner JL, Roche S, Kiel M, Toon GC, Wunch D, Baier BC, Biraud S, Chen H, Kivi R, Laemmel T, McKain K. A new algorithm to generate a priori trace gas profiles for the GGG2020 retrieval algorithm. Atmospheric Measurement Techniques Discussions. 2022 Oct 11;2022:1-41. https://doi.org/10.5194/amt-16-1121-2023, 2023

[7] Shen K, Tian Y, Hu B, Luo J, Qi S, Chen S, Lin H. Association rule mining of air quality through an improved Apriori algorithm: A case study in 244 Chinese cities. Transactions in GIS. 2024 Jun;28(4):726-45. https://doi.org/10.1111/tgis.13156

[8] Talukdar R, Sarma SD, Bostani A, Akbar S, Khamidova F, Mehbodniya A, Tarafdar T. Strategic Management of Technology and Organizational Innovation for Sustainability: A Policy-Oriented Analysis of Innovation Capabilities, Governance Mechanisms, and Long-Term Value Creation. Acta Innovations. 2026 Jan 6;59:1-8.

[9] Le M, Hoang DT, Nguyen DN, Pham QV, Hwang WJ. Wirelessly powered federated learning networks: Joint power transfer, data sensing, model training, and resource allocation. IEEE internet of things journal. 2023 Oct 13;11(21):34093-107. https://doi.org/10.1109/JIOT.2023.3324151

[10] Poli NS, Sikder AS. Predictive Analysis of Sales Using the Apriori Algorithm: A Comprehensive Study on Sales Forecasting and Business Strategies in the Retail Industry.: Predictive Analysis of Sales Using the Apriori Algorithm. International Journal of Imminent Science & Technology.. 2023 Nov 16;1(1):1-6. https://doi.org/10.70774/ijist.v1i1.1

[11] Tran DT, Huh JH. Forecast of seasonal consumption behavior of consumers and privacy-preserving data mining with new S-Apriori algorithm: DT Tran and JH Huh. The Journal of Supercomputing. 2023 Jul;79(11):12691-736. https://doi.org/10.1007/s11227-023-05105-6

[12] Egash D, Fatem BF. Digital Twin-Enabled Predictive Control of Electrical Machines and Converters. National Journal of Electrical Machines & Power Conversion. 2025 Oct 21:16-23. https://doi.org/10.17051/NJEMPC/01.03.03

[13] Alfaverh F, Denai M, Sun Y. A dynamic peer-to-peer electricity market model for a community microgrid with price-based demand response. IEEE Transactions on Smart Grid. 2023 Feb 17;14(5):3976-91. https://doi.org/10.1109/TSG.2023.3246083

[14] Kumar TS. Design and Performance Evaluation of a Bidirectional Power Electronic Interface for Renewable Energy Storage Integration. Transactions on Power Electronics and Renewable Energy Systems. 2025 Oct 16:25-32.

[15] Sun X, He Y, Wu D, Huang JZ. Survey of distributed computing frameworks for supporting big data analysis. Big Data Mining and Analytics. 2023 Jan 26;6(2):154-69. https://doi.org/10.26599/BDMA.2022.9020014

[16] Chen S, Xue Y, Cui X. Information literacy of college students from library education in smart classrooms: based on big data exploring data mining patterns using Apriori algorithm. Soft Computing. 2024 Feb 1;28(4):3571-89. https://doi.org/10.1007/s00500-023-09621-8

[17] Wang C, Niu Y, Zuo L, Yu R, Liu G. Decision-Making Method for County Power Grid Dispatching with High Proportion of Renewable Energy. Distributed Generation & Alternative Energy Journal. 2025 Sep 25:655-80. https://doi.org/10.13052/dgaej2156-3306.4042

[18] Wei Z, Zhang H, Zhang Y, Li B, Tao Y, Gao Y, Zhao C. Fast compressed wideband spectrum sensing. IEEE Transactions on Vehicular Technology. 2023 Sep 22;73(2):2924-9.

[19] https://doi.org/10.1109/TVT.2023.3318217

[20] Leyene T, Fahad AJ. AI-Enabled Internet of Energy Framework for Optimized Smart Grid Integration and Sustainable Renewable Energy Management. National Journal of Renewable Energy Systems and Innovation. 2025 Oct 21:1-8. https://doi.org/10.17051/NJRESI/01.03.01

[21] Uribe J, Shaik S. AI-Driven Energy-Efficient Electric Drive Systems for Renewable Energy and Industrial Automation Applications. National Journal of Electric Drives and Control Systems. 2025 Oct 21:10-6. https://doi.org/10.17051/NJEDCS/01.03.02

[22] Mira HR. Optimized Big Data Analytics Pipeline for Predictive Maintenance in Smart Manufacturing Systems. SECITS Journal of Scalable Distributed Computing and Pipeline Automation. 2024 Dec 5:15-23.

[23] Sindhu S. Scalable Data-Inclusive MLOps for Training and Operating Generative Models. Journal of Scalable Data Engineering and Intelligent Computing. 2025 Sep 27:26-36.

[24] Uvarajan KP. Fault-Resilient Coordination Algorithms for Uncertainty-Aware Wind Farm Micro-Siting in Distributed Networks. Transactions on Secure Communication Networks and Protocol Engineering. 2025 Sep 20:10-6.