Research on Grid Data Model Fusion Based on Big Data Extraction Technology

Zhao Ruifeng, Li Shiming, Liu Yang, Li Bo, Lu Jiangang

Electric Power Dispatching and Control Center of Guangdong Power Grid Co., Ltd., Guangzhou 510600, China

Abstract: The task of the electric power Internet of Things is to break through the various obstacles that form an open and shared energy ecology, which needs to be broken through by building an integrated energy system. Besides, the theoretical and practical application of cloud computing-based power big data task scheduling and extraction methods can help power companies to form the effect of resource sharing. Therefore, based on dynamic task priority, a scheduling algorithm is proposed in the paper to dynamically calculate the priority of all tasks, and effectively use idle periods to make tasks execute as quickly as possible to shorten the execution time, which is used for low-latency tasks such as power systems. After experimental analysis and comparison tests with other methods, it proves that the data extraction technology studied in the paper has the best effect and can be used to obtain better data sharing and access results in power companies.

1. Introduction
With the continuous growth of electric power data, the era of electric power big data has arrived, and the construction of smart grids is also gradually promoted. In terms of data processing, the demand for real-time and reliability is increasing. In recent years, with the rapid development of cloud computing and big data technologies, distributed processing platforms have emerged at the historic moment, which provides an infrastructure for the realization of big data processing. Moreover, for the power data center that is the foundation of information platform, as the amount of data continues to increase, requirements on its performance continue to increase. Therefore, research on cloud computing-based power big data scheduling and distribution methods has important practical significance[1-2].

Based on cloud computing and big data processing technology, a multi-queue dynamic priority scheduling algorithm and node selection algorithm for task scheduling and distribution are proposed in the paper, which is compared with the classic algorithm through simulation to verify its accuracy and effectiveness[3].

2. Scheduling and Distribution Algorithm

2.1 Task Scheduling Priorities

2.1.1 Task Value Feature Model
Since the data information is different, the effective value of the data processing task will be very different. Moreover, the data sources in the effective value include the importance $g$ and the data type $k$, and the data types are composed of regular type, abnormal type and alarm type. Additionally, the value of $g$ will affect the initial value and the maximum value of priority. Under the same conditions,
if $g$ value of key equipment data task in power system is high, the task will be processed in time. Moreover, $k$ reflects the urgency of the task. Under the same conditions, the time-sensitive data priority changes quickly, so the priority is high as well. In addition, due to the influence of $g$ and $k$ values, the highest priority and time spent for each task are also different[4-5]. Equation (1) is the effect of the two parts in effective value on the total value.

$$W_i = a g_i + b k_i$$  \hspace{1cm} (1)$$

In formula (1), $W_i$ is the effective value of processing task, $a$ and $b$ are coefficients, and $g_i$ refers to the key value of data source. Besides, $k_i$ is the value of data type.

### 2.1.2 Task Time Feature Model

In smart grid environment, there are more business source systems and more requirements for processing time and speed. Therefore, it is necessary to assign priorities based on time requirements to ensure that the timeliness of data is arranged by size.

The time characteristic is also an important factor that affects the priority. In addition to the value of $g$ and $k$, it can also reduce blocking and improve data processing efficiency if the allocation of waiting time and processing tasks is reasonably adjusted.

Most power monitoring data contains time attributes, as shown in equation (2).

$$\{a_i, b_i, d_i, s_i, t_i, w_i\}, \quad i < n$$  \hspace{1cm} (2)$$

where $a_i, b_i, d_i, s_i, t_i$ and $w_i$ are the start execution time, actual completion time, deadline time, arrival time, expected execution time and waiting time. In addition, the time characteristics of the data are determined by the effective value and the waiting time. The effective time is $d_i-s_i$. Equation (3) shows the impact of each attribute on the total value[6-7].

$$T_i = \frac{\lambda w_i}{d_i - s_i}$$  \hspace{1cm} (3)$$

In formula (3), $T_i$ is the time characteristic of the $i^{th}$ processing task, and $\lambda$ refers to the coefficient.

### 2.1.3 Dynamic Priority Assignment Strategy

The priority of most scheduling algorithms is sorted by task processing time, and few consider their value. Power data is not only limited by time, but also has a significant impact on its effective value. Therefore, a dynamic priority allocation strategy is proposed. When a large amount of data arrives at the same time, especially in special cases such as faults and alarms, the scheduling optimization will be completed according to the priority. Otherwise, the data may be blocked and lost. Equation (4) is the assignment of the dynamic priority in the task.

$$P_i = \alpha T_i + \beta W_i$$  \hspace{1cm} (4)$$

In formula (4), through a large amount of training, the time feature $\alpha$ and the effective value weight coefficient $\beta$ are obtained from the corresponding data set. What’s more, when entering large-scale data, some data processing tasks cannot be completed in time due to certain factors, which still have important value and will not be discarded within a certain period of time after the deadline. However, for critical data such as warning data, the priority is always kept at the maximum, even it exceeds the limit[8-9].

In other data, after the deadline, the priority decreases with time, and its change is shown in equation (5).

$$P_i = P_{\max} - \frac{z_i - d_i}{w_i}(T_{\max} + k_i)$$  \hspace{1cm} (5)$$

In Equation (5), $P_{\max}$ is the maximum priority value of the task, $z_i$ refers to the current time;, and $T_{\max}$ indicates the maximum priority value occupied by the time characteristic part of the task.
2.2 Task Distribution Mechanism

The task distribution mechanism is mainly to ensure the reliability of data and avoid loss, and the distribution mechanism is based on hierarchical cloud computing cluster model. According to the different requirements of the task, it is sent to various control centers in line with factors such as time, priority, and data backup is distributed to the surrounding nodes to avoid node loss and data loss. The distribution mechanism is as follows:

1) When the cluster nodes analyze the data and find the alarm data, the data will be encapsulated to send to the distribution queue, then the data will be broadcast to all nodes in the area. Besides, the alarm data will be sent to the superior control node;
2) If abnormal data is found, the regional central control node will select multiple nodes for data backup, which will not be processed immediately;
3) The routine data collection is performed in a queue order, and tasks that require shorter time can be scheduled in layers.

3. Experimental Results and Analysis

3.1 Simulation Parameters

The master-slave structure is composed of 2 sub-areas and 2 central nodes, and each area consists of 1 area control node and 5 slave nodes. What's more, Storm and Kafka serve as the foundation of large-scale data cloud platforms, and Storm relies on zookeeper to ensure and maintain cluster consistency. Additionally, commands can be run directly on a single node. If a multi-node cluster needs to be configured, config/server.properties parameters must be configured. In order to verify the effectiveness of the algorithm, it is compared with the classic algorithms EDF and DVD in scheduling and distribution experiments.

3.2 Comparison on Completion Rate of Important Tasks

As shown in Figure 1, there are the experimental results of the three algorithms scheduling in the cluster, and the completion rate of important tasks is compared.

![Fig.1 Comparison of completion rate of important tasks](image)

It can be seen from Figure 1 that EDF algorithm has the highest completion rate at the initial stage. As the task grows, although it is scheduled and distributed to surrounding nodes, there are still many tasks to be processed, which prioritize tasks based only on deadlines. As deadlines approach, regular tasks will preempt computing resources to complete analysis and processing.

Compared with EDF algorithm, the algorithm proposed in the paper and DVD algorithm can complete most tasks, but the initial completion rate is lower. As the number of processing tasks increases, DVD algorithm will have infinite priority since some tasks are close to the deadline, which leads to decrease in completion rate. However, as the calculation of average density is a priority, the completion rate of key tasks is higher than that of EDF algorithm. In addition, the algorithm proposed in the paper also considers the impact of time and value on processing tasks, and the priority can be
dynamically adjusted according to the cutoff time. However, due to its value, the priority will not be infinite to preempt resources. Therefore, the completion rate gradually decreases, and the larger the number of tasks is, the higher the completion rate will be.

### 3.3 Comparison of Load Average Dispersion

With the increase of the number of task data, the variation of load dispersion during the scheduling of each algorithm is shown in Figure 2.

![Fig.2 Load average deviation comparison](image)

It can be seen from Figure 2 that due to the influence of unilateral factors, EDF has a higher load in cluster scheduling. Besides, DVD algorithm takes into account the changes in value and time, so it is more efficient in cluster scheduling, and the load balance difference is smaller than that of EDF algorithm. What’s more, the algorithm proposed in the paper not only considers the impact of time and value, but also avoids the possibility of infinite priority, which can effectively distribute the load and reduce the load difference between the receiving node and other nodes.

Experimental results show that the cluster hierarchical model and both algorithms have good performance. Moreover, from the task completion time, key task completion rate and load average dispersion, the algorithm proposed in the paper significantly improves the task completion rate and completion time, and its performance is significantly better than other algorithms.

### 4. Conclusion

In terms of cloud computing and big data technology, a multi-queue dynamic priority scheduling algorithm and node selection algorithm for task scheduling and distribution is proposed in the paper, and the simulation is used to verify the rationality of the algorithm. It is proved that method proposed in the paper has obvious advantages in task completion rate and time. In addition, due to current laboratory hardware requirements and the scale of experimental data, research on data dispatch and distribution methods for smart grids based on big data is still in primary stage. Therefore, gradually improving and perfecting on the algorithm will become the focus of the next work.

### References

[1] H.W. Hui, C.C. Zhou, S.G. Xu, F.H. Lin, A Novel Secure Data Transmission Scheme in Industrial Internet of Things, China Communications, vol. 17, no. 1, 2020, pp: 73-88.
[2] F.H. Lin, Y.T. Zhou, X.S. An, I.You, K.R. Choo, Fair Resource Allocation in an Intrusion-Detection System for Edge Computing: Ensuring the Security of Internet of Things Devices, in IEEE Consumer Electronics Magazine, vol. 7, no. 6, 2018, pp: 45-50. doi: 10.1109/MCE.2018.2851723.
[3] J.T. Su, F.H. Lin, X.W. Zhou, X. Lv, Steiner tree based optimal resource caching scheme in fog computing, China Communications, vol. 12, no.8, 2015, pp: 161-168.
[4] Jiang Dong, Liu Kunpeng. Research on double-layer real-time optimization game of active distribution network [J]. Electric Measurement & Instrumentation. 2018 (18)
[5] Ren Jinxia, Huang Yipei, Zhong Xiaokang. Improved Cloud Task Scheduling Algorithm Based on Genetic Algorithm [J]. Journal of Jiangxi University of Science and Technology. 2018 (03)

[6] Huang Wenjing, Li Huaiqiang, Yang Zhiya, et al. Risk assessment of power grid security based on fuzzy clustering ranking and state uniformity [J]. Electric Measurement & Instrumentation. 2018 (09)

[7] Liu Scientific Research, Sheng Wanxing, Zhang Dongxia, et al. Research on Big Data Application Requirement and Scenario Analysis of Intelligent Distribution Networks [J]. Proceedings of the CSEE, 2015 (02)

[8] Peng Liping, Lv Xiaodan, Jiang Chaohui, et al. Elastic scheduling strategy for cloud resources based on Docker [J]. Journal of Computer Applications. 2018 (02)

[9] Wang Yueyue, Chen Minura, Jiang Zhenchao, et al. State assessment of secondary equipment in intelligent substations based on cloud theory [J]. Power System Protection and Control. 2018 (01)