Research on Power Big Data Processing Attribute Reduction Method Based on Cloud Computing Technology

Hui Zhang[1, a], Fande Kong[2, *] and Haiwen Wang[1, b]

[1]School of Economics and Trade, Ji Lin Engineering Normal University, Changchun 130000, China
[2]Qinhuangdao Highway Engineering Quality Supervision Station, Qinhuangdao 066000, China

*Corresponding author Email: 214528699@qq.com, [a]187004499 @qq.com, [b]39414698@qq.com

Abstract. Cloud computing data processing and calculation makes the processing of power big data intelligent and fast. The cloud computing technology will be complex, large, multi-power data actualized and pooled. The cloud computing data processing center becomes the current power big data. Important technology. Based on the background of cloud computing technology, this paper proposes a power big data Processing reduction, and gives a brief overview of the theory. Combined with the MapReduce model, the paper implements the power big data monthly processing reduction application. Finally, the attribute reduction of the grid fault diagnosis table and the wind power measured data is carried out on the Hadoop platform. The experimental results show that the method is effective and feasible, and has a good speedup and applicability. It is suitable for the power big data Processing attribute. Simple.

1. Introduction

The power industry is the basic supporting industry of the national economy. The development and application of power industry information and power production automation have generated massive amounts of data in power companies. With the application of technologies such as Internet of Things, cloud computing, and e-commerce in the power industry, the data volume of power enterprise data centers has rapidly increased from several hundred terabytes to several thousand terabytes [1]. The data growth rate is getting faster and faster, and the power industry is entering the era of big data. Power big data has the universal characteristics of large data, such as large amount of data, multiple types, rapid changes, and high value. It is of great practical significance to dig deep into the value of power big data for power enterprise management, power production and social energy conservation [2].

Attribute reduction can reduce data dimensions, reduce unnecessary storage and irrelevant input, and significantly improve the efficiency of power data Processing. Faced with the massive increase in massive power data, computers in power systems are facing bottlenecks in storage and computing resources. It is difficult to meet the rapidly growing demand by simply improving the level of hardware and software. As a new generation of parallel programming system, MapReduce utilizes its

unique elastic distributed data set MP_POSRS to process large-scale power data sets in parallel on the basis of existing software and hardware. This paper takes the massive short-term power forecast data of a wind farm as an example, and introduces MapReduce into the knowledge reduction algorithm. Since most of the data in the power prediction table is continuous, the continuous attribute must be discretion. Due to the characteristics of the knowledge splitting and demonetization technology, some information may be lost [3]. In order to ensure the integrity of knowledge, the author deeply studied the MapReduce programming model, analyzed the relative positive domain theory of rough sets and the existing knowledge reduction algorithm, and used the properties of relative positive domain to give the precondition of power big data Processing. The related definitions and theorems of Simplified Chinese and the MapReduce model are used to design the MP_POSRS algorithm for parallel computing the relative positive potential of the power big data set. The Hadoop platform is used to implement the power big data Processing attribute reduction algorithm in the cloud environment. The experimental results show that the proposed algorithm not only can efficiently calculate the attribute reduction of power big data sets, but also has good applicability.

2. Related definitions and theorems of power knowledge expression systems

Hadoop is an open source implementation of the Map-Reduce parallel programming framework proposed by Google. The MapReduce program consists of a Map function and a Reduces function. Each time the Map function converts an input (key, value) pair into a set of intermediate (key, value) pairs; the Reduce function processes the same set of values for the key, resulting in a final the result is written to the distributed file system HDFS. MapReduce, a relatively efficient parallel programming model, can solve the problem of power big data Processing attribute reduction. The following are the more representative definitions and theorems [4].

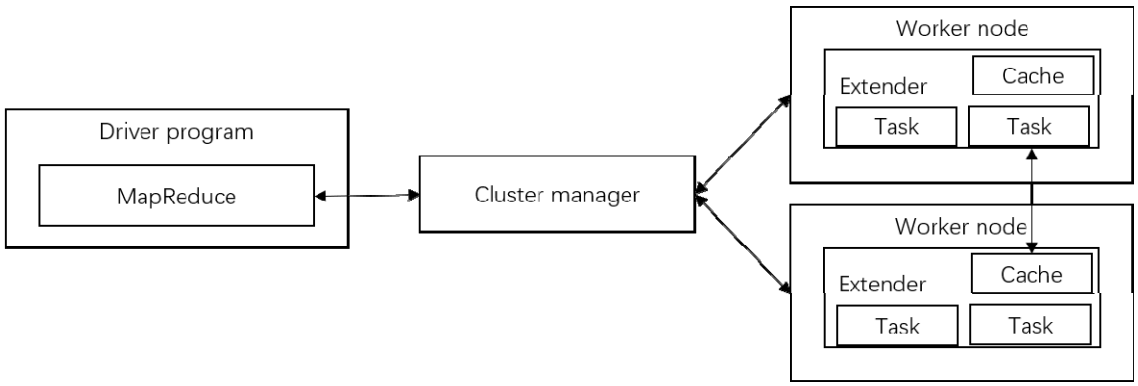

Fig 1. MapReduce running architecture diagram

2.1. Definition

Definition 1: Assume that the power knowledge expression system is the decision information table S, $S=(U,A,f,V)$, U is the object set $A=C\cup D$, C is the condition attribute set, D is the decision attribute set, V is the attribute value set, f is the information function, and the object is clear Attribute value [5].

Definition 2: Let $P,Q\in A$, $P\cap Q=\varnothing$, P be the positive domain of Q, and $pos_p(Q)=\bigcup_{x\in U/Q}\underline{}^{PX}$, $Count\left(pos_p(Q)\right)$ denote the number of elements contained in P.

2.2. Theorem

Theorem 1: Assume that the power knowledge expression system $S=(U,A,f,V)$, $P,Q\in A$, $P\cap Q=\varnothing$, $R\subseteq P$, $Count\left(pos_R(Q)\right)=Count\left(pos_P(Q)\right)$ is a necessary and sufficient condition for $pos_R(Q)=pos_P(Q)$; the necessity proves that: because $pos_R(Q)=pos_P(Q)$, the P positive domain of Q is the same as the positive R of Q, $Count\left(pos_R(Q)\right)=Count\left(pos_P(Q)\right)$; proof of sufficient : Proof by the counter-evidence method, that is, let $pos_R(Q)=pos_P(Q)$ not hold, and because $R\subseteq P$, so $pos_R(Q)\subseteq pos_P(Q)$, and because $pos_R(Q)=pos_P(Q)$ does not hold, then $Count\left(pos_R(Q)\right)<Count\left(pos_P(Q)\right)$, and $Count\left(pos_R(Q)\right)=Count\left(pos_P(Q)\right)$ contradict, so it is not established [6].

Theorem 2: Let $S=(U,A,f,V)$ be a power knowledge expression system, $P,Q\in A$, $A=C\bigcup D,C\bigcap D=\varnothing$, C be the conditional attribute set, D be the decision attribute set, and $a\in C$ is the necessary and sufficient condition of the necessary attribute. Proof of necessity: $Count\left(pos_{C-\{a\}}(D)\right)\neq Count\left(pos_C(D)\right)$ is a nuclear attribute, then $pos_{C-\{a\}}(D)\neq pos_C(D)$ is defined by definition 1, and $Count\left(pos_{C-\{a\}}(D)\right)\neq Count\left(pos_C(D)\right)$ is known by theorem 1. Proof of sufficiency: Because $Count\left(pos_{C-\{a\}}(D)\right)\neq Count\left(pos_C(D)\right)$, we know $pos_{C-\{a\}}(D)\neq pos_C(D)$ from Theorem; we know that a is a nuclear property.

The above is some of the definitions and theorems in the MapReduce parallel programming model. It can be seen that the use of rough set theory in the process of attribute reduction of a power knowledge representation system can effectively reduce the complexity of key attribute reduction, which can not only be effective. Reducing the amount of calculation of the entire reduction process can also better reduce the consumption of time and resources, while the power big data Processing attribute reduction based on cloud computing technology is based on the premise of strengthening the cloud the application of computing technology to further improve its efficiency [7].

3. MapReduce programming design of reduction algorithm

Considering a power big data set as a power knowledge expression system, the corresponding conditional attribute of the specified decision attribute set is required, that is, the attribute reduction problem of this power big data set is converted into the calculation positive field. The problem of the situation [8]. MapReduce is used to calculate the above problem. The specific method is as follows: the map function accesses multiple data fragments at the same time, and takes out the attributes and attribute values according to actual needs, and generates <key, value> key-value pairs. The meaning of the representative is <"CO11 fault area Sec1", 1>). The Reduce function receives a sequence of key-value pairs corresponding to the key values sent from the maps of the respective nodes, and uses this to determine the specific number of the same equivalence class.

(1) The map function is located in the same time period for each of the multiple data fragments to independently expand access, and at the same time according to the actual requirements of the specification to obtain the attribute and attribute values, and then generate the key value pair <key, value>; (2) Reduce function pair That is, the sequence of key-value pairs corresponding to the map value sent by each node at the node is also calculated and processed for the corresponding number of equivalence classes. When Hadoop is applied to complex tasks, it focuses on increasing the number of tasks, not on the complexity of map and Reduce functions. Therefore, in the cloud data-based power big data Processing attribute reduction, two maps, three Reduce and call-job functions are designed,

and a master program can be carried at the same time, and finally combined with actual needs. Each given algorithm, the reduction calculation can be performed for the big data Processing attribute [9].

4. Experimental analysis

Traditional knowledge reduction methods cannot handle large data sets, so this section does not compare with traditional methods, and discusses its application in power big data Processing only from the influence of the number of nodes.

This article uses Hadoop platform to build a cluster experimental environment consisting of 16 laptops, of which Hadoop version is Hadoop-0.20.0, the maximum configuration of notebook computer is dual core 2.50GHz, 4GB memory, 1TB hard disk, the lowest Configured as dual-core 2.00GHz, 1.5GB RAM, and 160GB hard drive. The experimental data is wind power measured data, including 14 attributes, and the size is 24GB. Realize the extraction, transformation and loading (ETL) of experimental data, fill the gap value with the average value of adjacent data, discreet the data into a series of 0,1 list, to improve the data processing efficiency, and then get 13 Conditional attribute, one decision attribute of power knowledge representation system S [10]-[12].

4.1. Applicability

Applicability is the performance of parallel algorithms when the data size is scaled up by the number of nodes. In order to test the applicability of the algorithm, four samples of 2.5, 5, 10, and 20 GB were taken from the experimental data as test data sets, and scale and aging experiments were performed on 2, 4, 8, and 16 nodes, respectively. As shown in Figure 2 [13]. It can be seen from the figure that although the performance of the algorithm decreases slightly when the number of nodes increases to 16 due to hardware and platform operation resource consumption, the running time of these operations basically maintains the same level, which shows that the parallel algorithm of this paper is good. Applicability.

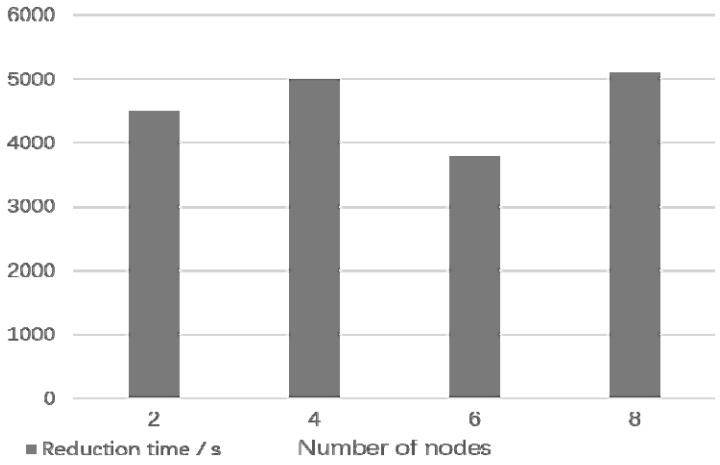

Fig 2. Applicability test

4.2. Acceleration ratio

The speedup ratio is the performance of the parallel algorithm when the data size is fixed and the number of nodes is continuously increased. The ideal speedup is linear, but due to the overhead of communication between computers, task scheduling, etc., the actual speedup will be lower than ideal. The test data set size is 20GB, and the number of nodes is 2, 4, 8, and 16, respectively. As shown in Fig. 3, it can be seen from the relationship between the reduction time and the number of nodes in the figure that the parallel algorithm of this paper obtains good acceleration performance [14].

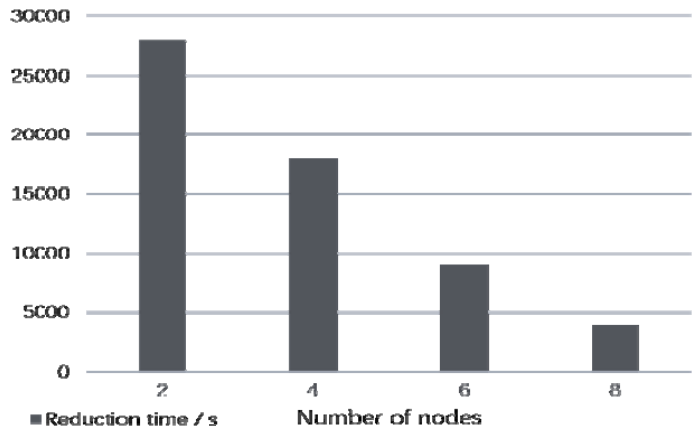

Fig 3. Acceleration ratio test

5. Conclusion

With the rapid development of the smart grid construction process, the data collection of the power system has grown geometrically, that is, stepping into the era of power big data. The traditional non-naturalized heuristic attribute reduction algorithm encounters challenges in processing power big data. Although the improved parallel heuristic attribute reduction algorithm overcomes this bottleneck, it is inherent in the heuristic attribute reduction algorithm. In the process of reduction, some attributes with low attribute importance are lost, which causes partial loss of information in the decision table. The partial order reduction algorithm studied in this paper not only solves the problem of decision table information loss caused by heuristic attribute reduction algorithm, but also skips the nucleation process of heuristic attribute reduction algorithm, using departmentalization of partial order method. The feature is applied to the MapReduce framework to directly reduce the related attributes of the power data. Finally, by simulating on the Hadoop platform, the results show that the partial order method can be used for the reduction of power big data, and the time performance of the algorithm is good.

References

[1] Wu Kaifeng, Liu Wantao, Li Yanhu, et al. Power Big Data Analysis Technology and Application Based on Cloud Computing. China Electric Power, Vol.2 (2015) No.48, p. 111-116.
[2] Wang Wei, Yu Xiuli, Liu Xiaojun. Power Big Data Analysis Technology and Application Based on Cloud Computing. Mobile Information, Vol.3 (2015) No.12, p. 00079-00084
[3] Mao Dong, Yan Xubin, Shen Zhihao, et al. Research on power big data attribute reduction method “Electronic Technology Application” Smart Grid Conference. Vol.5 (2017) No.14, p.80-85.
[4] Pi Yulin. A Simple Method for Processing Attributes of Power Big Data Based on Cloud Computing Technology. Science and Technology Innovation Review, Vol.12 (2017) No.14, p. 158-159.
[5] Yu Yu. Research on key technologies based on parallel heuristic reduction method. North China Electric Power University, Vol.1 (2015) No.16, p. 70-72.
[6] Xu Feifei, Lei Jingsheng, Bi Zhongqin, et al. Global Approximation Reduction of Interval Values for Multiple Decision Tables in Big Data Environment. Journal of Software, Vol.9 (2014) No.12, p. 2119-2135.
[7] Wang Xinxin. demonetization method of power big data attributes based on cloud computing technology. Digital Technology and Applications, Vol.1 (2015) No.32, p. 56-58.
[8] Chen Qi. Research on Power Big Data Feature Analysis Based on Hadoop. North China Electric Power University (Beijing), Vol.3 (2016) No.15, p. 172-175.

[9] Liang Jiye, Qu Kaishe, Xu Zongben. Attribute Reduction of Information Systems. Systems Engineering - Theory & Practice, Vol.12 (2001) No.21, p. 76-80.
[10] Xu Feifei, Lei Jingsheng, Bi Zhongqin, et al. Global Approximate Reduction of Interval Values for Multiple Decision Tables in Big Data Environment. Journal of Software, Vol.12 (2014) No.9, p. 2119-2135.
[11] Zhang Ziqian, Zhang Yanyan, Chen Wei, et al. Research on 3D Visualization Management Method of Power Big Data. Energy and Environmental Protection, 2018(3). Vol.3 (2018) No.27, p.159-163.
[12] Song Yu, Jiao Ji, Li Gang. Analysis of the Characteristics of Attribute Reduction in Big Data Processing. Journal of Computer Measurement and Control, Vol.12 (2015) No.23, p. 4191-4194.
[13] Zhang Yun, Xu Ning. Definition and Reduction of Attribute Importance of Big Data Sets Based on Equivalence Classes. Chinese Journal of Scientific Instrument, Vol.14 (2004) No.13, p. 801-803.
[14] Yu Wei. Research on Key Technologies Based on Parallel Heuristic Reduction Method. North China Electric Power University, Vol.1 (2015) No.11, p. 35-39.