Mathematical Model of Data Partition Storage in Network Center Based on Blockchain

Bing-bing Han[1], Zai-xing Su[1], and Hai-yun Han[2(✉)]

[1] Panjin Vocational and Technical College, Panjin 124000, China
dnsdsdf2l@aliyun.com
[2] School of Humanities and Social Sciences, Sanya Aviation & Tourism College, Sanya 572000, China
sdgsahdl232@aliyun.com

Abstract. In view of the poor effect of data partition storage in current network center, which affects the effect of data storage, this paper proposes a mathematical model of data partition storage in Network Center Based on blockchain. Firstly, the cross attribute storage algorithm based on blockchain is designed. Through the algorithm, the function of network center data is analyzed, and different functions are analyzed in detail. On this basis, the method of network center data division is designed. Finally, the network center data is divided into two parts. Build the mathematical model of data partition storage in network center to improve the cache performance. The experimental results show that the proposed method can better meet the requirements of the central data partition storage mathematical model cache operation efficiency, and provide better storage and operation performance.

Keywords: Blockchain · Network center · Central data · Data storage · Mathematical model

1 Introduction

Today's Internet applications pay more and more attention to the access efficiency of data partition storage in the network center. In order to better improve the effect of data storage, we must maximize the use of processor for partition processing, and store the data that may be used in multi-level buffer [1].

Based on this, this paper proposes the construction of the mathematical model of network center data partition storage based on blockchain, so as to better in the process of partition storage, the location of data placement is particularly important for the optimization of cache utilization. Therefore, in the operation of data partition, we should choose a good data storage scheme, improve the spatial layout of data distribution, improve the utilization of cache, and improve the performance. The storage layout of each attribute value in the record is adjusted, and some attributes in the record are accessed according to the requirements, so as to eliminate the memory delay caused by unnecessary memory access.

S. Wang et al. (Eds.): IoTCare 2021, LNICST 414, pp. 187–199, 2022.
https://doi.org/10.1007/978-3-030-94185-7_13

2 Mathematical Model of Data Partition Storage in Network Center

2.1 Data Feature Management in Network Center

With the rapid increase of data volume, network center data feature management has been updating. Static data is the core of data partition center in Network Center, and the feature of static data is metadata, so the collection of metadata is a key. In the aspect of metadata collection, the most traditional way of collection management is to store the program and data together, which brings many restrictions to the flexibility of the system, so many data and programs have begun to collect and store separately [2]. Until the emergence of database feature management information, the separation of metadata and basic data is realized. Metadata management is mainly in two aspects. One is the storage of metadata, which serves as the access docking and storage place of metadata. The second is a way of metadata exchange. Currently, there are three main metadata management models, such as Table 1:

Table 1. Data management model of Network Center

Type	Explain	Advantage	Shortcoming
Centralized model	There is a metadata server that serves the metadata store with the client request	Simple implementation	Single point of failure, limited storage
Distributed model	With the cooperation of multiple hosts, it serves as the metadata management center of the whole system	Solve some single point problems	Metadata synchronization is expensive and difficult to design
No metadata model	Using hash algorithm to replace metadata service in distributed system	Eliminate performance bottlenecks, single point of failure, etc.	Increase data management complexity

In the aspect of distributed database, the distributed database can be regarded as a whole in the logical layer and a distributed data system in the physical layer. Distributed database has unique advantages in applicability, reliability, availability and scalability. Before the tide of Internet and informatization changes the demand of database, we are faced with a small number of users. The traditional relational database can fully cope with it, and the distributed database is put aside because of its consistency problem [3]. However, in the face of a huge amount of data, the traditional relational database technology has been unable to meet the requirements of massive data. As a result, the requirement of data consistency is lowered, and NoSQL database, which is often used now, appears. Column storage and traditional row storage are two corresponding storage methods [4]. In many practical scenarios, what is needed is only an individual field of a record, and too many queries are unnecessary, especially in the massive data environment, the cost can not be underestimated [5]. For the applicable

column storage, the query is grouped according to the "column", and the query is concentrated on the corresponding column, which makes the operation of individual attributes more convenient. The current mainstream column storage database is shown in the Table 2.

Table 2. Data sequence of network partition

Mainstream representatives	Hbase, Cassandra, Hypertable
Data model	Centralized data storage by column
Advantage	For column query, IO has obvious advantages and strong scalability
Shortcoming	Data integrity is no better than row storage

2.2 Data Partition Method in Network Center

Data storage is related to the use of data, performance, and management difficulty. Data quality evaluation is related to user experience, resource allocation, and accuracy of results. Data storage and quality evaluation are very important for the efficient use of many systems [6]. There are many ways to store and organize data. The hardware level has the storage mode set for the speed difference between the cache and the hard disk, and the software level stores the data separately according to the reading and writing, so as to reduce the pressure of the database. In fact, the development of more common is the sub table and partition storage technology. Sub table and partition are similar, the purpose is to improve the operation efficiency, both are very similar, both need rule decomposition table [7]. The difference is that table splitting divides a large table into multiple entity tables, while partition is for data segments. There are many ways to evaluate the data. Generally speaking, it is mainly in the following four aspects: the consistency of the data, whether the data information is accurate enough, whether the data is complete and whether the data is excessive, and the timeliness is poor [8]. Based on this, the distributed data and traditional database are analyzed and designed. Distributed storage will combine the data characteristics of blockchain and select mature NoSQL database as the basic storage. Blockchain is designed for massive data registration, including static data and dynamic data. Static data is the information registered in the form of metadata to manage the registration data. Facing the storage bottleneck of massive data environment, both static data and dynamic data must adopt distributed storage [9]. The data partition repository structure of the network center is as follows (Fig. 1):

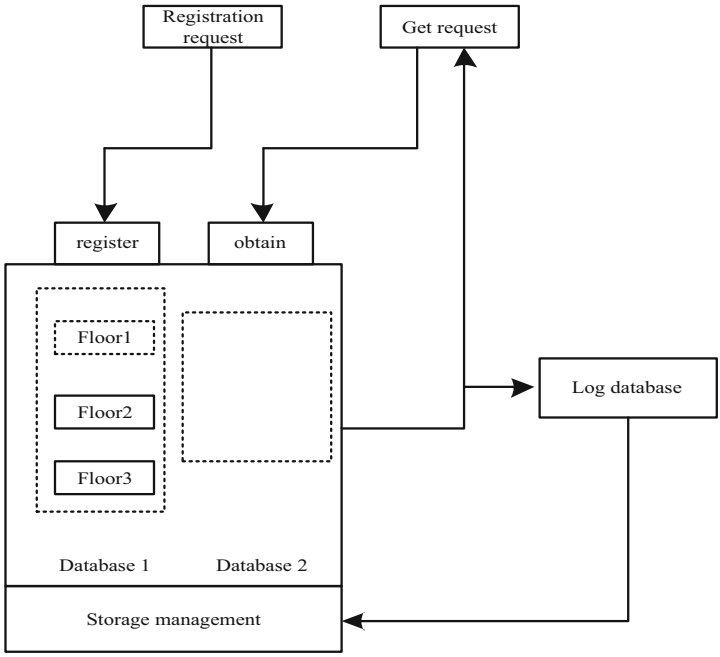

Fig. 1. Data partition repository structure of Network Center

Data consistency algorithm for different scenarios, the distributed consistency of mass data is mainly the consistency of multiple copies. The current methods are also relatively diverse. The consistency methods for distributed systems mainly have solutions, mainly with the help of transaction control method to achieve the effect, as well as the extended transaction control method based on the extension, through the data replication The replication control method for controlling, the message queue method for controlling messages, and the primary replica update message queue method. It is self-evident that the accuracy of data is only the basic standard of information availability, and the distorted data itself can be regarded as the data without quality. Most of the researches in this field are in the state of model and lack of uniform applicability. From the perspective of decision maker, the timeliness of data is related to the accuracy of decision. From the perspective of ordinary users, timeliness is related to the selectivity of users. From the perspective of system, timeliness is related to the offset of system overhead. Complete the data balancing partition, and its basic composition is shown in the figure below (Fig. 2).

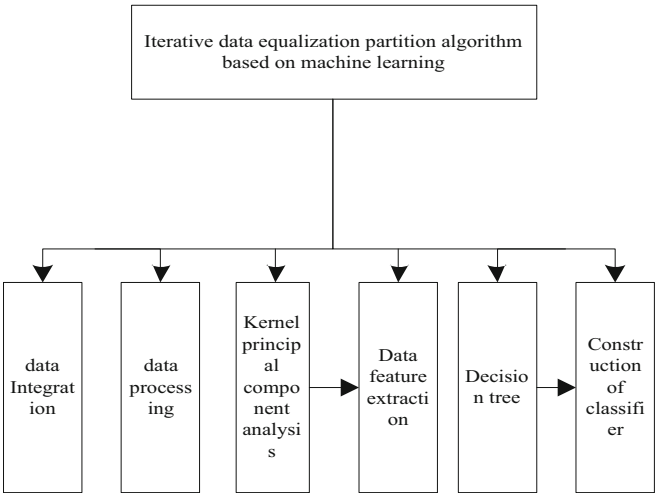

Fig. 2. Composition of data balanced partition structure

It can be seen from the figure that the iterative data equalization partition algorithm based on machine learning includes three parts.

In the first part, heterogeneous data is integrated and processed to eliminate invalid data and reduce data scale.

The second part is to extract data features.

In the third part, the classifier model is constructed to realize the data Iterative balanced partition.

As a logical data pool, blockchain is the storage place for data registration and storage. As shown in the figure, as the core of data oriented architecture, blockchain is responsible for internal data management and external interaction to provide access and location services. "Physical" cloud refers to the data cloud of the physical world, such

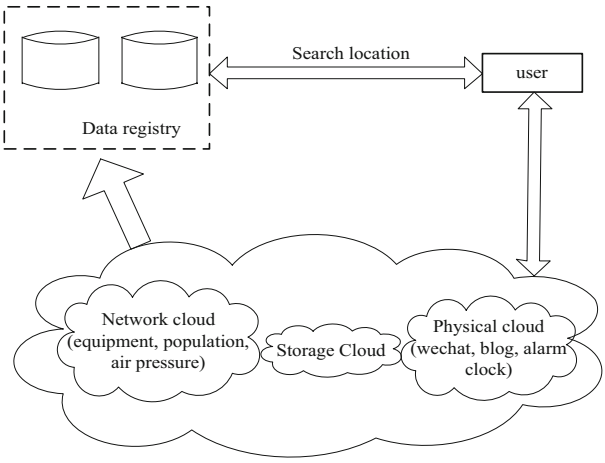

Fig. 3. Data partition processing logic of Network Center

as all kinds of natural substances and all kinds of entity attributes; “network” cloud refers to all kinds of attribute data cloud in the network (Fig. 3).

It further shows that the dynamic data forms a chain structure according to the time. Here, the static data has two operation records of “dynamic data 1” and “dynamic data 2”, which are arranged in turn. It can also be seen that the dynamic data tracking is the static data information, and the design is to present the chain structure in turn according to time. The unique identifier is used as the link relationship identifier. With the help of dynamic data, we can fully obtain the change process of static registration information, and track the operation change of data according to time. This will play an auxiliary role in the authority supervision of the later authority components. The relationship between dynamic data is shown in the Fig. 4.

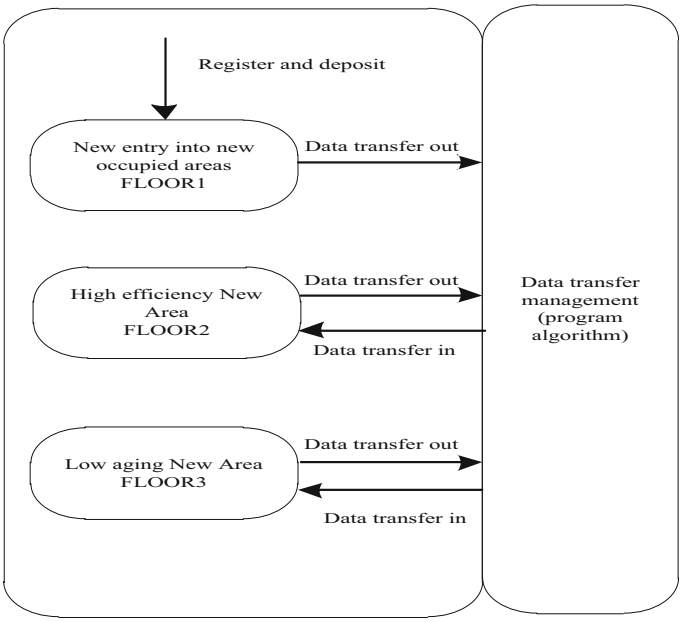

Fig. 4. Dynamic data relationship

In view of the rapid growth of data, as a distributed database, blockchain adopts automatic fragmentation mechanism to realize the distributed storage of massive data. In this paper, we will analyze the fragmentation mechanism of the blockchain, and design the patch key based on the characteristics of the blockchain itself and the wide range of data access statistics. At the end of data integration and processing, and as a sample data, data feature extraction. Based on principal component analysis, the kernel method is applied to form the kernel principal component analysis method. A set of data to be analyzed is mapped into a suitable high-dimensional feature space according to a nonlinear mapping rule, and then the data is processed and analyzed in this space by using a linear learner

$$O(x_i, y_i) = \sum_{n=1}^{k} \zeta(x_i, y_i)n \tag{1}$$

In the formula, $O(x_i, y_i)$ is the coordinate of the data in the high-dimensional feature space after nonlinear mapping; k is the space dimension; n is the number of data in the data sample; ζ is the nonlinear mapping rule, that is, the mapping function.

Assuming that the order of the curve filled in the massive data space is M, the S space of the massive data set can be divided into $2^M \times 2^M$ grids, and each grid has a four-dimensional spatial Hilbert code:

$$M_0 = O(x_i, y_i)H\left[\log_2 \frac{D_0}{H_1}\right]2^M \times 2^M \tag{2}$$

In the formula: D_0 represents the total amount of data; H_1 represents the storage size of data block. Statistical coding data element information set O, assuming that the total number of data coding blocks is i, if the data block storage size H is greater than the maximum threshold percentage of massive data blocks, then the coding fast is divided into sample set y_i.

According to the aggregation characteristics of spatial four-dimensional Hilbert coding linear filling curve, the massive data coding blocks are decomposed, and the corresponding storage sequence of each coding block H_2 is marked to form the corresponding spatial data partition matrix, as follows:

$$F = \begin{vmatrix} H_{2\text{de}0} & H_{1a0} & s_0 \\ \cdots & \cdots & \cdots \\ H_{2\text{de}n} & H_{1an} & s_n \end{vmatrix} \tag{3}$$

In the formula: $H_{2\text{de}n}, H_{1an}, s_n$ represents the spatial element of massive data, according to which the corresponding spatial code and the corresponding massive data block storage label after matrix matching are obtained, thus the spatial element division of massive data is completed.

According to one of the items, the corresponding massive data set is divided. Assuming that the length is larger than the width, then the horizontal axis direction of massive data is the positive direction set, and the vertical axis direction is the opposite direction set, so the spatial element interval of massive data is calculated [10]. Through this interval, the coding block is decomposed to realize the feature partition of massive data. Combined with the organization layer technology of relational database, the three-stage aging model is designed. The design focuses on the reasonable distribution of data, the design of data aging algorithm, and the three-tier design of storage area according to the aging [11]. In the face of large amount of data and high throughput conditions, there are two basic methods to solve the problem of insufficient compression performance: vertical expansion and fragmentation. Monodb adopts fragmentation mechanism.

The fragmentation mechanism of blockchain is shown in the figure data fragmentation [12]. Data is divided into multiple places to share the storage, which is the

basis of massive data storage in the system. There are mainly two ways of slicing: by range and by hash (Fig. 5).

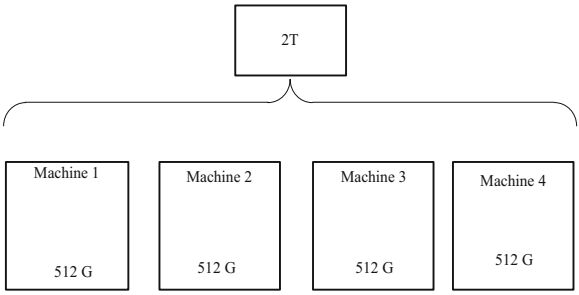

Fig. 5. Data fragmentation processing method

Storage control is responsible for data transfer among them. The three layers are new data area, high aging data area and low aging data area.

2.3 Implementation of Data Partition Storage

The data source is not necessarily uniform, and the format is not necessarily the same, which is not conducive to partitioning and reduces the speed of partitioning, so it is necessary to integrate these heterogeneous data before partitioning. Data integration is the process of integrating data of different sources, formats and characteristics into one big data, so as to serve the subsequent partition. Here, data integration is mainly completed by central data warehouse technology, and its structure is shown in the figure below (Fig. 6).

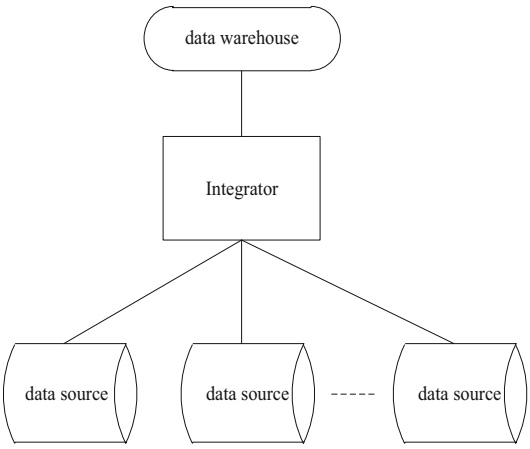

Fig. 6. Data partition processing library

The biggest advantage of central data warehouse is that it can achieve the maximum control of the extracted data, and its application effectively solves the problems of data dispersion, diversity and redundancy.

After data integration, further data processing is needed, including data cleaning and data reduction.

Data cleaning is to find and correct abnormal data in data set, including data consistency check, invalid value and missing value processing, duplicate data deletion, etc.

Data reduction refers to reducing the amount of data and reducing the scale of data. The amount of data after integration is huge. If it is directly used in the follow-up work, it will not only increase the amount of data, but also reduce the accuracy of partition. Therefore, it is necessary to reduce the original data after integration, including data aggregation, dimension reduction, data compression, data block reduction, etc.

The timeliness of data is an important aspect of data value. In the past, the judgment of data timeliness was mainly based on the fact that the data was only valid in a certain period of time. But for the generalized data, it is not enough to only consider this point. There are not many algorithms to determine the timeliness of data, but the design is to consider the efficiency of data query and the cost of system resources. The three-stage partition setting of the system has different overhead configurations for different partitions. In order to improve the adaptability and controllability of the system, the following time effective algorithm is designed on the basis of the original "one shot dead" and "dominated by the user", so as to achieve the function of controllable allocation of index resources. In the design of timeliness judgment algorithm, the controllability of timeliness is mainly considered.

$$N = M_0 i + 3TF \tag{4}$$

$$\mathrm{f}(x) = a_0 + \sum_{N=i}^{\sum} (f_1(c_i, c_{i+1})^* f_2(c_i, c_{i+1})) \tag{5}$$

$$f_1(c_i, c_{i+1}) = \mathrm{sgn}(c_{i+1} - c_i) \tag{6}$$

$$f_2(c_i, c_{i+1}) = \frac{|c_i - c_{i+1}|}{\max(i_i, c_{i+1})} \tag{7}$$

In the formula: a_0 is the reference value, as the floating reference. c_i here is the number of visits in the T cycle.

The number of visits in four consecutive t-cycle time in turn. T here for a day or more. The longer it is, the easier it is to cover the downlink line, and the slower it is to respond to the short-term steep drop in the number of visits. The segmentation value of high and low timeliness is L. The aging value above L is high aging, and below L is low aging. Set the high efficiency value as a. Above a value is high aging, below a value is low aging.

$$\begin{cases} \operatorname{sgn}(f(x) - A) = 1 \\ \operatorname{sgn}(f(x) - A) = -1 \end{cases} \tag{8}$$

For the first new data area, there is a regular transfer out, the amount of data is relatively stable, there will be no data accumulation. This area is indexed on keywords according to the query. The source of data in this area is registration. Transfer out includes two directions, one is the second layer of efficient data area, the other is the third layer of inefficient data area. The direction of transfer out depends on the storage management module. According to a judgment calculation result of aging algorithm, if the timeliness of T cycle is good, it will enter the high aging region, otherwise it will enter the low aging region. For the efficient data area of the second area, the data of this area is also in and out, and the volume is relatively stable. The worst case is that all data is often used, resulting in data accumulation here. According to the data characteristics of the three regions, the data resources are allocated. Index is very important for data query performance, and it costs a lot. When the index size exceeds the memory capacity of the system, it will bring too much V0 performance consumption. So there is a tendency in the configuration (Table 3).

Table 3. Resource allocation

Hierarchy	Configuration policy
FLOOR1	Configure keyword index, title index and subject
FLOOR2	Configure keyword index, title index and subject
FLOOR3	Configure title index

Further optimize the data partition storage state, as shown in the Fig. 7:

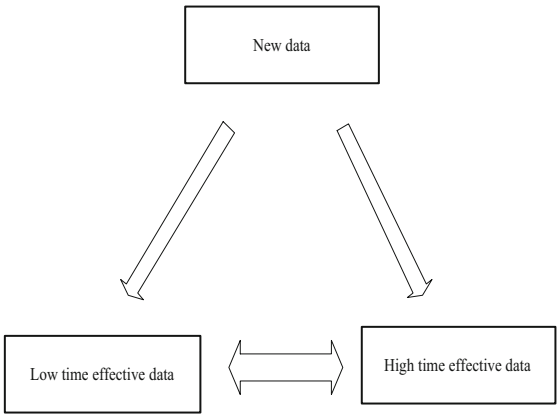

Fig. 7. Data partition storage status

The status of data can be divided into new data, high aging and low aging. The area the data should belong to is determined by its aging value. Different from the other two areas, new data can only be registered in one data entry mode. According to the access situation in a certain period of time, there are two destinations. The data of the two outer regions flow through each other, and the data from the new data area can be obtained. The timeliness of data is an important aspect of data value. In the past, the judgment of data timeliness was mainly based on the fact that the data was only valid in a certain period of time. But for the generalized data, it is not enough to only consider this point. Not all the data.

3 Analysis of Experimental Results

In order to test the effectiveness of data balancing partition algorithm, simulation test is needed. First of all, the test needs to clarify the experimental environment, as shown in the following Table 4.

Table 4. Simulation test environment

Name	Parameter setting
Operating system	Windows 764 bit
CPU	Intel, corei5-24110m memory 4G
Application server	Ggda-25t5 special server
Server database	Oracle 10g
Simulation experiment software	Matlab R2009a
Test recording tool	Dev test tool

The storage balance test scheme is shown in the Table 5.

Table 5. Storage balance test scheme

Test items	Storage balance test
Test purpose	Test whether the data put in is balanced
Prefabrication conditions for testing	1. Subject and_ ID is the partition key
	2._ ID system insertion. Subject randomly generates character strings
	3. Each block is limited to 1m
	4. Set two servers as two partitions
Testing procedure	1. Take FL2 as the object, insert 10000000 pieces of metadata
	2. Check the blocking condition
Expected results	The data block distribution of the two partitions should be relatively balanced

In order to test the data equalization ability, the segmentation limit of each data block is 1m. Insert 10 million pieces of data, data storage balance display shard0000, shard0001 each contains 796 and 797 data blocks. In the experiment, the size of each block is limited to 1m, and the amount of data stored in each chip is basically the same. The test scheme is shown in the table. The time-consuming statistics of the measured results of the network center data partition test scheme are shown in the Fig. 8.

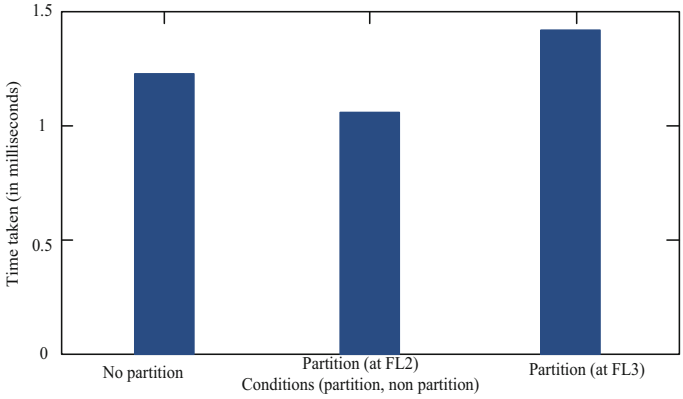

Fig. 8. Comparative analysis of partition storage time consumption

The basis of data partition in network center is that the active data is a small part, which can save time when querying a small range. It is proved that the block chain is more effective in the network center data partition storage model.

4 Conclusion

Under the background of big data era, the development of network information technology makes it possible that all kinds of information is no longer an information island. In view of the rapid growth of data, how to form a data oriented development system with data as the core is of great significance to data sharing and management. Based on this, combined with the blockchain to build the network center data partition storage model, in order to better deal with massive data effectively. Although the mathematical model constructed in this paper can effectively deal with network center data, there are still many shortcomings. In the future, we will study the changing attributes of network center data and the increasing amount of data to improve the performance of this method.

References

1. Li, X., Jiang, P., Chen, T., et al.: A survey on the security of blockchain systems - ScienceDirect. Future Gener. Comput. Syst.**107**(4), 841–853 (2020)

2. Fraga-Lamas, P., Fernandez-Carames, T.M.: Fake News, disinformation, and deepfakes: leveraging distributed ledger technologies and blockchain to combat digital deception and counterfeit reality. IT Prof. **22**(2), 53–59 (2020)
3. Gao, X., Ren, B., Zhang, H., et al.: An ensemble imbalanced classification method based on model dynamic selection driven by data partition hybrid sampling. Expert Syst. Appl. **160** (5), 113–115 (2020)
4. Nomura, K., Shiraishi, Y., Mohri, M., et al.: Secure association rule mining on vertically partitioned data using private-set intersection. IEEE Access 1 (2020)
5. Levitin, G., Xing, L., Huang, H.Z.: Security of separated data in cloud systems with competing attack detection and data theft processes. Risk Anal. **39**(4), 846–858 (2019)
6. Nozal, R., Perez, B., Luis Bosque, J., et al.: Load balancing in a heterogeneous world: CPU-Xeon Phi co-execution of data-parallel kernels. J. Supercomput. **75**(3), 1123–1136 (2019)
7. Nguyen, T.D., Lee, S.W.: PB-NVM: a high performance partitioned buffer on NVDIMM. J. Syst. Arch. **97**(11), 20–33 (2019)
8. Shanthi, P.A., et al.: Privacy preserving time efficient access control aware keyword search over encrypted data on cloud storage. Wirel. Pers. Commun. **109**(4), 2133–2145 (2019)
9. Zhang, Y.F., Wang, X.P., Pan, Y.X., et al.: Alteration in isotopic composition of gross rainfall as it is being partitioned into throughfall and stemflow by xerophytic shrub canopies within water-limited arid desert ecosystems. Sci. Total Environ. **692**(20), 631–639 (2019)
10. Liu, S., Liu, D., Srivastava, G., Połap, D., Woźniak, M.: Overview and methods of correlation filter algorithms in object tracking. Compl. Intell. Syst. **7**(4), 1895–1917 (2020). https://doi.org/10.1007/s40747-020-00161-4
11. Liu, S., Lu, M., Li, H., et al.: Prediction of gene expression patterns with generalized linear regression model. Front. Genet. **10**, 120 (2019)
12. Fu, W., Liu, S., Srivastava, G.: Optimization of big data scheduling in social networks. Entropy **21**(9), 902–908 (2019)