Academic Journal of Engineering and Technology Science
ISSN 2616-5767 Vol.4, Issue 2: 1-6, DOI: 10.25236/AJETS.2021.040201

Research on Smart City Double-link Data Storage and Sharing Model based on Alliance Chain

Luo Wenmin[*]**, Jin Jiangtao**

Institute of Big Data and Internet Innovation, Hunan University of Technology and Business, Changsha 410205, China
[*]*Corresponding Author*

Abstract: *Most of the existing sharing models adopt blockchain technology. With the increase of data volume and participating nodes, the storage performance and sharing security of data cannot be guaranteed, which is prone to data leakage. To solve this problem, we propose a double-link data storage and sharing model based on the alliance chain, which combines off-chain data storage and on-chain data sharing. The off-chain database uses the Hadoop Distributed File System (HDFS) distributed file system based on mean shift clustering to store data, which improves the storage and access rate of the data. The consensus mechanism combining Delegated Proof of Stake (DPOS)and Practical Byzantine Fault Tolerance (PBFT) with copy deletion scheme is adopted to realize the safe sharing of data. The proposed model has good efficiency, security, and accuracy, and can effectively improve the problems existing in smart city data security management.*

Keywords: *smart city; alliance chain; information security; data storage; data sharing*

1. Introduction

With the construction and development of smart cities, the amount of global data generated every year has jumped to the ZB level, and the scale of data generated every year continues to expand. With the rapid increase of data volume and data types, the security management needs of smart city urban data cannot be met by traditional data storage and data sharing methods. The privacy of thousands of college students in many places has been leaked, and many students have been employed; There are frequent leaks of private data, and data security has become a core element of smart city construction. Data sharing is the key to the realization of smart cities. Alliance chain is a kind of block chain, which has the characteristics of decentralization, privacy, fast transaction processing, and strong privacy protection. It can solve the problem of scattered information resources and information security of smart cities. As the number of participating nodes and stored data increases, the computational overhead of the alliance chain becomes larger and larger, and the storage space required gradually increases, which results in the performance of data storage and sharing decreases, thus affecting the scalability of the alliance chain. Therefore, we propose a double-link data storage and sharing model based on the alliance chain. The off-chain database uses the HDFS distributed file system based on mean shift clustering to store data to ensure data storage and access speed. The alliance chain adopts the DPOS+PBFT consensus mechanism with the copy deletion scheme to solve the problem of too many participating nodes, which leads to too long consensus. The consensus mechanism has certain fault tolerance, which can effectively avoid malicious node attacks and realize the safe sharing of data.

2. Related research

Big data is faced with the problems of difficult storage and low reading rate. HDFS system has the advantages of processing large files and storing heterogeneous data, which has become the key to solve the problem of data storage in smart cities. Ranjitha et al. improved HDFS memory utilization by implementing new efficient deduplication technology [1]. He Long et al. proposed the general index technology and two-level index mechanism to improve query efficiency [2]. Xie Guojun et al proposed a dynamic decentralized storage optimization strategy CDDS based on Cauchy code to reduce the required storage space and improve the system reliability and load balancing ability [3]. In the existing studies, methods such as removing redundant data and adjusting load balancing are adopted to improve the efficiency of data storage, and the double-level index mechanism is used to filter invalid access and reduce the time of data reading from disk. However, the problem of low data search efficiency still exists when effective access is conducted. Therefore, on the basis of removing redundant data and filtering invalid access requests, we use the mean shift algorithm to build a multi-level data storage and index structure to improve the efficiency of effective access operations.

With the deepening of research on secure data sharing, the role of blockchain in secure sharing is gradually highlighted. Zhang et al. [4], Ge Lin et al. [5], Tan Haibo et al. [6] realize shared data privacy protection by using blind signature algorithm, smart contract, digital signature, and other technologies. Zhang Chao et al. designed an alliance Medical block Chain system based on PBFT algorithm, to solve problems such as easy tampering and leakage of Medical data [7]. Wang Tong et al. proposed the framework of information sharing and secure multi-party computing model based on block chain, which reduced computing cost and increased data transmission to a certain extent [8]. The alliance chain has the characteristics of strong controllability, semi-centralization, and data will not be disclosed by default. It is the most promising blockchain. The Practical Byzantine Fault Tolerance Algorithm (PBFT), as a common consensus mechanism for alliance chains, allows consensus to be reached in the presence of malicious interference from a small number of nodes and increases the fault tolerance of the blockchain. However, consensus efficiency depends on the number of participating nodes, so it is not suitable for blockchain with too many nodes. In addition, PBFT algorithm cannot store transaction information well, and hackers will intercept part of invalid copies, which will lead to information leakage. In order to solve the above problems, we adopt the consensus mechanism combining DPOS and PBFT, and add the effective copy deletion mechanism, so as to improve the efficiency of data consensus and ensure the security of data sharing.

3. Double - link data storage and sharing model framework based on alliance chain

In order to realize efficient storage and safe sharing of data, the double-linked data storage and sharing model framework based on alliance chain is constructed, as shown in Figure 1.

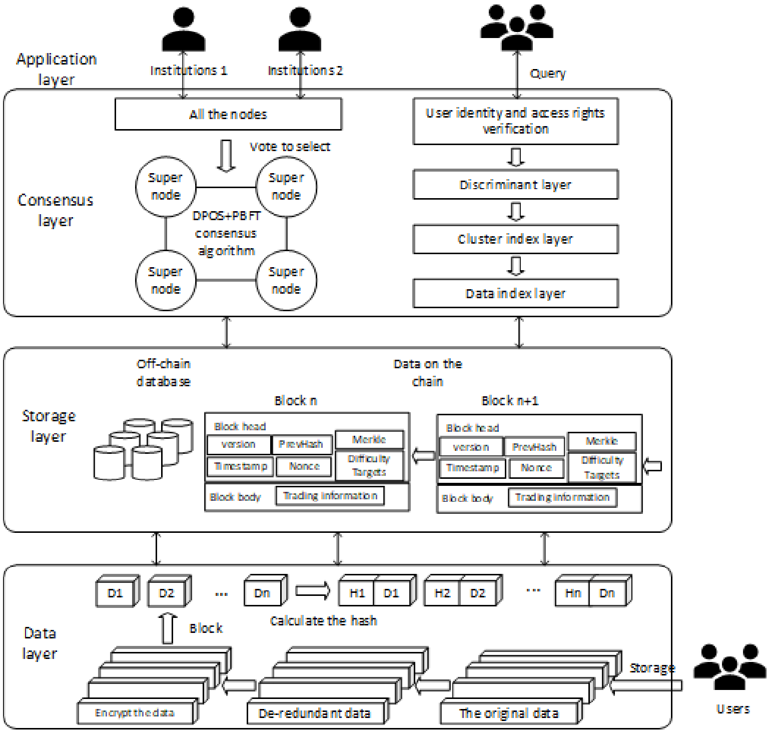


Figure. 1 The framework of double-linked data storage and sharing model based on alliance chain

In the data layer, redundant data cleaning, data encryption and data partition are carried out on the uploaded data by users, and the hash value of each data block is calculated to ensure data security.

In the storage layer, considering that with the rapid increase of data volume, the management difficulty of the alliance chain will be increased, we adopt the double-chain storage mode. The off-chain database stores encrypted data information. Data description, data storage address and other related index information are stored on the chain. According to the data characteristics, the index data is stored by means of mean-shift clustering. On the one hand, the double-chain mode for data storage releases a large amount of storage space on the alliance chain and improves the sharing efficiency. On the other hand, the mean-shift clustering method for data storage is helpful to improve the data access speed.

In the consensus layer, considering the increasing number of nodes added in the alliance chain, the workload of consensus will be huge. Therefore, the consensus is made by the combination of DPOS and PBFT. Super nodes are selected by voting, and then the super node makes consensus on the new data block and data access request, effectively improving the consensus efficiency.

In the information indexing, the simple data index structure is changed into a multi-layer index structure including the discriminant layer, the cluster index layer, and the data index layer. The discriminant layer can reduce The Times of invalid access and improve the efficiency of other users to access data. The cluster index layer identifies the cluster where the data is located according to the characteristics of the data accessed, and then searches the data storage address from the data index layer. This index structure further improves the data access speed.

In the application layer, when the organization needs to access and obtain data information, it needs to send a query request to the alliance chain, and the super node verifies the user's identity and access authority. After successful verification, it can search the encrypted data information in the database under the chain through the data address in the index information.

4. Detailed design of the model

The model uses alliance chain technology to efficiently store and securely share data information. The model is divided into two parts: off-chain data storage and on-chain data sharing. Double-chain storage is adopted, a large amount of data is stored in an off-chain database, and only index information such as data descriptions and data addresses are stored in the alliance chain. According to the characteristics of mass and heterogeneity of smart city data, the distributed file system based on mean shift clustering is adopted to store the data in the off-chain database to improve the efficiency of data access. The consensus mechanism of DPOS+PBFT is adopted to improve the consensus speed, and then improve the efficiency of on-chain data storage and sharing.

4.1 Off-chain data storage model

Due to the large amount of redundancy in the collected data, direct storage will waste a large amount of storage space. In addition, big data contains various types of unstructured data, which increases the difficulty of data storage and reading. Therefore, after using the hash function to delete redundant data, the HDFS distributed file system based on mean shift clustering is used to improve data access efficiency, which is of great significance for big data storage.

After obtaining the data, the data is divided into blocks, the MD5+SHA1 digital signature algorithm is used to calculate the hash value of each data block, and then the hash value is searched in the hash index table. If the same hash value is found, it is duplicate the data block and delete; otherwise, consider it as a new data block, store the new data block and add the hash value of the new data block to the hash index table. Repeat the above steps to clean up redundant data. This method can effectively improve the utilization of storage space and can also further improve storage performance and storage efficiency.

Data blocks are stored in multiple servers, corresponding nodes are generated and extended to multiple file systems, and a file system network is composed of many nodes to realize unstructured data storage. The client is connected with the node through the computer network. When the client makes a request to access the data, it will automatically look for the server of the data storage and extract the data block according to the node. The client can access the data without knowing the location of the data storage.

HDFS uses a streaming data access method, which has better performance when accessing all data, but there will be a problem of high latency when random access to data is required. In order to solve this problem, a multi-layer index HDFS scheme based on mean shift clustering is proposed. Through this clustering algorithm, the data index structure composed of NameNode is transformed into a multi-layer discriminant layer, cluster index layer, and data index layer. Layer structure, this multi-layer index structure reduces data access latency. The mean shift clustering algorithm can automatically determine the number of clusters to complete the clustering according to the characteristics of the stored data, so it is suitable for situations where the number of clusters cannot be determined due to different stored data. The steps to apply this solution to achieve big data storage are as follows:

1) Perform redundant cleaning through the cleaning process in Figure 2 to obtain a de-redundant data set.

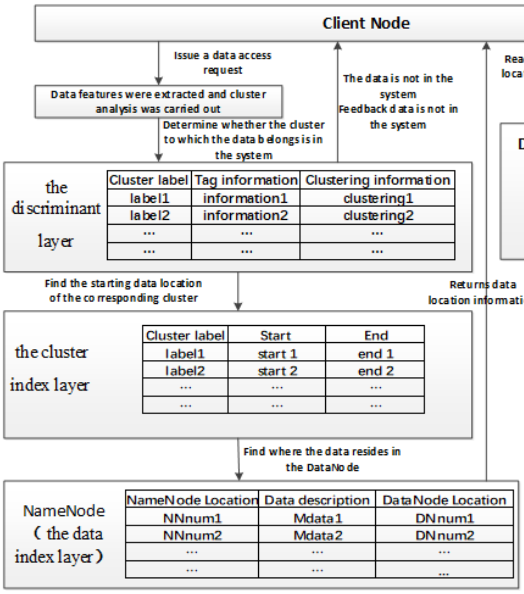
2) Extract original data features, use clustering algorithm, and generate different clusters based on data features.

3) According to different clusters, store the data characteristics and data location information in the NameNode as the data index layer.

4) Generate the corresponding cluster label, all cluster labels, the start and end positions in the cluster constitute the cluster index layer.

5) Make all cluster labels, label descriptions, and cluster information into a statistical table to form an access discrimination layer.

6) When users need to add stored data, first go through redundancy cleaning to obtain de-redundant data, extract data characteristics, determine clusters, add data to clusters, and update the statistical table of the discrimination layer.


Client Node
Issue a data access request
Data features were extracted and cluster analysis was carried out
Determine whether the cluster to which the data belongs is in the system
The data is not in the system
Feedback data is not in the system
Read data based on location information
Send the correspon ding location data
DataNode
the discriminant layer
Find the starting data location of the co rresponding cluster
Returns data location information
the cluster index layer
Find where the data resides in the DataNo de
NameNode （the data index layer）

Cluster label	Tag information	Clustering information
label1	information1	clustering1
label2	information2	clustering2
...	...	...
...	...	...

Location	Data
DNnum1	data1
DNnum2	data2
...	...
...	...

Cluster label	Start	End
label1	start 1	end 1
label2	start 2	end 2
...	...	...
...	...	...

NameNode Location	Data description	DataNode Location
NNnum1	Mdata1	DNnum1
NNnum2	Mdata2	DNnum2
...	...	...
...	...	...

Figure. 2 Data access flow chart

When a user needs to access a certain data, the data features will be analyzed and judged at the discriminant layer. If there is no cluster label matching with it, the data information will be fed back to the user; otherwise, the cluster label matching with the data features will be found and the data will be searched in the cluster (as shown in Figure 2). In the case of random access to a certain data, the discrimination layer discriminates the invalid access request of the user, that is, the data accessed by the user is not in the system, By reducing the amount of disk access, improving the efficiency of other users' access to data. The clustering layer determines the cluster of data by analyzing the data characteristics.

4.2 On-chain data sharing model

When using alliance chain technology for data sharing, a combination of on-chain and off-chain storage mode is adopted. Data location information is stored on the chain, and data is stored in HDFS distributed file system based on mean shift clustering under the chain. Data storage and access of the alliance chain are carried out through consensus mechanism. Consensus mechanism is the key to realize decentralization and ensure safe data sharing. In the process of consensus, many participating nodes lead to long transaction time and slow transaction speed. The combination of Byzantine fault-tolerant algorithm and non-Byzantine fault-tolerant algorithm can effectively improve consensus efficiency. Because the Byzantine algorithm cannot store the copy information well, the hacker intercepts the invalid copy, which will lead to data leakage. Therefore, on the basis of improving the consensus efficiency by using the consensus mechanism of DPOS and PBFT, the copy deletion scheme is added to further guarantee the security of data sharing.

The improved consensus algorithm contains n consensus nodes and provides the fault tolerance of f=(2n+1)/3, where f is the maximum number of malicious nodes allowed to exist. According to the DPOS mechanism, r nodes are selected as super nodes and the other nodes are ordinary nodes. The process of data sharing in the alliance chain is as follows.

1) After the user submits the transaction request, the master node is calculated according to the height of the block, the number of transaction requests and the number of super nodes.

2) The user makes a transaction request to the local node with the public key as the identity.

3) If the local node is not the master node, the request will be sent to the next node until the master node is found and the request will be verified. If the user has no authority, the message will be fed back to the user, return to step 1, recalculate the master node, and start processing the next transaction request; If the user has permissions, then consensus begins.

4) The user uses the Hash function to get the data summary from the original data, uses the private key to digitally sign the data summary, uses the public key to encrypt the data, timestamp and encrypted digital signature, and sends the encrypted information to the master node.

5) The master node will broadcast the copy information and the inspection results of permissions to other super nodes.

6) The super node receives a copy of the information for review and broadcasts the review results. After receiving the review results of other super nodes and comparing them with their own review results. If the review results are inconsistent, the response results will be sent to the master node, and the response results will be a comparative analysis of the review results of the node itself and those of other super nodes. When any super node receives (2n+1)/3 different nodes all approve the validity and security verification of the current transaction, it is deemed to have reached a consensus and the verification ends.

7) The master node analyzes the review results and response results, determines whether the nodes have malicious behaviors, and deals with the malicious nodes in time to ensure the safe operation of the system. The data is packaged to generate blocks and sent to all nodes. Send block and delete transaction record orders to all super nodes.

8) All nodes update the alliance chain, super node deletes the transaction record and copy information, and feedback the response results of block update and copy deletion to the master node.

9) If it exceeds the consensus time or fails to receive (2n +1)/3 different nodes' verification pass messages, the verification will be considered as failure, and the failure information will be fed back to the client. The broadcast stops verifying and deleting the transaction record information. After receiving the information, the super node stops verifying, deleting the transaction record information and the copy information, and feedback the response result.

10) Recalculate the master node and start the next round of trading consensus.

5. Conclusion

In the context of big data, the development of smart cities cannot be separated from scientific decision support. It is increasingly important to fully tap the value of big data to assist decision making. With the increase of data capacity, data type and transmission speed, new requirements are constantly put forward for data management methods. However, the existing data management methods are difficult to give full play to the value of big data. In order to meet the development needs of smart city data security sharing, we based on the double-link data storage and sharing model framework of alliance chain, to deal with the challenges faced by the security storage and sharing of big data with complex structures.

References

[1] S.Ranjitha, P.Sudhakar and K.S.Seetharaman (2016). "A Novel and Efficient De-duplication System for HDFS". Procedia Computer Science, vol.92: p.498-505.
[2] L.He, J.C.Chen and X.Y.Du (2017). Multi-Layered Index for HDFS-Based Systems. Journal of Software, vol.28, no.03, p.502-513.
[3] G.J.Xie, J.Q.Shen and H.H.Yang (2019). An HDFS storage optimization strategy based on Cauchy code. Computer Engineering & Science, vol.41,no.03, p.440-445.
[4] P.Zhang, W.Jules, C.S.Douglas ,L.Gunther and R.S.Trent (2018). "FHIRChain: applying blockchain to securely and scalarly share clinical data."[J]. Computational and structural biotechnology journal vol.16, p. 267-278.
[5] L.Ge, X.S.Ji, T.Jiang and Y.M.Jiang (2019). Security mechanism for Internet of things information sharing based on blockchain technology . Journal of Computer Applications, vol.39, no.02, p.458-463.
[6] H.B.Tan, T.Zhou, H.Zhao, Z.Zhe,W.D.Wang, Z.X.Zhang, N.Z.Sheng and X.F.Li (2019). Archival Data Protection and Sharing Method Based on Blockchain. Journal of Software, vol.30, no.09, p.2620-2635.
[7] C.Zang, Q.Li, Z.H.Chen, Z.R.Li, Z.Zhang (2019). Medical Chain: Alliance Medical Blockchain System. ACTA AUTOMATICA SINICA, vol.45, no.08, p.1495-1510.
[8] T.Wang, W.P.Ma, W.Luo (2019). Information Sharing and Secure Multi-party Computing Model Based on Blockchain. Computer Science, vol.46, no.09, p.162-168.