7-23-2019

Design of Personnel Big Data Management System Based on Blockchain

Houbing Song
Embry-Riddle Aeronautical University, songh4@erau.edu

Jian Chen
Qingdao University

Zhihan Lv
Qingdao University

Scholarly Commons Citation

Song, H., Chen, J., & Lv, Z. (2019). Design of Personnel Big Data Management System Based on Blockchain. *Future Generation Computer systems, 101*(). https://doi.org/10.1016/j.future.2019.07.037

Future Generation Computer Systems 101 (2019) 1122–1129

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Design of personnel big data management system based on blockchain

Jian Chen [a], Zhihan Lv [a,*], Houbing Song [b]

[a] Qingdao University, China
[b] Embry-Riddle Aeronautical University, USA

HIGHLIGHTS

- A method for solving blockchain information redundancy is proposed.
- Analyze the shortcomings of the blockchain and the reasons for it.
- Feasibility of combining blockchain with information management system.
- Developed a prototype system to validate our ideas.

ARTICLE INFO

Article history:
Received 22 May 2019
Received in revised form 24 June 2019
Accepted 16 July 2019
Available online 23 July 2019

Keywords:
Big data
Blockchain
Information management
Data separation

ABSTRACT

With the continuous development of information technology, enterprises, universities and governments are constantly stepping up the construction of electronic personnel information management system. The information of hundreds of thousands or even millions of people's information are collected and stored into the system. So much information provides the cornerstone for the development of big data, if such data is tampered with or leaked, it will cause irreparable serious damage. However, in recent years, electronic archives have exposed a series of problems such as information leakage, information tampering, and information loss, which has made the reform of personnel information management more and more urgent. The unique characteristics of the blockchain, such as non-tampering and traceability make it have great application potential in personnel information management, and can effectively solve many problems of traditional file management. However, the blockchain is limited by its own shortcomings such as small storage space and slow synchronization time, and cannot be directly applied to the big data field. This paper proposes a personnel management system based on blockchain, we analyzed the defects of the blockchain and proposed an improved method, constructs a novel data storage model of on-chain and out-of-chain that can effectively solve the problem of data redundancy and insufficient storage space. Based on this, we developed a prototype system with query, add, modify, and track personnel information, verified the feasibility of applying blockchain to personnel information management, explore the possibility of combining blockchain with big data.

1. Introduction

Personnel information management has always been an indispensable part of human society's life and work. It includes file management of company and government employees, school students and teachers, as well as registered members of hotels and airlines, even included the national credit information system. How to store such huge and private information securely has become a problem for many companies and governments [1]. Now the mainstream personnel management system is using the B/S architecture to centrally manage all personnel information. In this architecture, the user client can change the information stored in the central database at any time after obtaining the license. The central administrator can control the database, has high privileges, and can authorize other users to access or even modify the database [2]. The biggest drawback of this architecture is that it centralizes the storage of data, and someone has the highest authority to operate on that data. The risk of data leakage and tampering is very high. For example, in 2017, the US Pentagon exposed the US Department of Defense database, which contains personal information collected by the United States on the global social media platform of 1.8 billion users. and Yahoo announced in 2016 that more than 3 billion account information was stolen. In 2014, China's largest online ticketing website 12306 was attacked by hackers, causing hundreds of thousands of citizen information to be leaked [3].

* Corresponding author.
E-mail address: lvzhihan@gmail.com (Z. Lv).

https://doi.org/10.1016/j.future.2019.07.037
0167-739X/

Because the personnel information management system inevitably stores a large amount of private information, therefore, the disclosure of such information directly leads to the safety of the person being stored in the database, if a leak accident occurs, it cannot be saved at all. If information such as ID card or telephone number is leaked, it is impossible to ask so many people to modify their ID cards or telephone numbers to avoid the risk of information leakage [4]. Secondly, such information management systems as credit reporting systems and academic systems involve the interests of many people, so the risk of being tampered with is very high. The disadvantage of a centralized database is that if someone has the right to modify the information, any changes can be made to the information, although the modification log is saved, but it is still saved in the centralized database and can still be deleted and modified [5]. Therefore, the difficulty in personnel information management in the field of big data is how to ensure the security of information is not leaky, and cannot be tampered and traceable.

The current mainstream method of central database to deal with these problems is to improve the difficulty of obtaining data management authority, to improve the security of access control, but it still depends on whether the decision is correct and trustworthy, and whether the decision center is safe [6]. Blockchain is a distributed database system based on peer-to-peer network, it is the result of integrating many technologies, these technologies include P2P protocol, zero-knowledge proof, consensus mechanism, smart contract, this creates a new way of storing and processing data differently than before [7,8].

The blockchain is composed of several data blocks linked together according to the order of generation time, and the data block can be generated through the consensus mechanism of each node, and the security is ensured by the encryption technology, so if a node tampers with a block, it is impossible to write the block to the entire system through the consensus mechanism [9]. According to the hash value, Merkel tree and time stamp can trace the operational history of each block [10].

Blockchain has the characteristics of decentralization, non-tamperability and programmability, which can effectively solve the security problem of big data storage, especially for the protection of personnel information, which involves a large amount of private information, and need to be regulated and endorsed by government and large companies [11–13].

But the blockchain itself has some serious drawbacks, the most important of which is its limited storage space, which makes it impossible to store large amounts of data. So if we want to use the blockchain to solve the security problem of big data storage, we must also solve the problem of limited storage space in the blockchain.

2. System architecture of personnel information management based on blockchain

Blockchains are divided into three categories according to the admission mechanism, Public Blockchain, Private Blockchain and Consortium Blockchain [14]. The Public Blockchain is the earliest and most widely used blockchain. Bitcoin is a representative Public Blockchain. Its characteristics are complete decentralization and not regulated or controlled by any institution, anyone can participate in the Public Blockchain [15]; The Private Blockchain is a system that is not open to the outside world and is used only within the organization [16]; the Consortium Blockchain is between the Public Blockchain and the Private Blockchain, it is usually used in the fields where multiple user roles such as companies, governments, and banks exist simultaneously [17].

The personnel information management system involves a lot of private and sensitive information. The Public Blockchain allows nodes to join freely, so it cannot be developed using the Public Blockchain [18]. However, for example, citizen information can be shared by the government, banks, and universities to a certain extent, therefore, its characteristics are in line with the requirements of Consortium Blockchain, so we use the Consortium Blockchain as the basic blockchain architecture of the personnel information management system [19]. Hyperledger Fabric is a blockchain framework implementation and one of the Hyperledger projects hosted by The Linux Foundation, Intended as a foundation for developing applications or solutions with a modular architecture, Hyperledger Fabric allows components, such as consensus and membership services, to be plug and play [20,21]. Hyperledger Fabric is a leading open source and universal blockchain structure for companies. Its throughput can reach 2000 transactions per second (TPS). Currently, there are more than 250 companies and organizations using it, including IT giants such as IBM, Intel, Baidu, Huawei, and other financial institutions such as ABN Amro, Accenture, and ANZ [22–24].

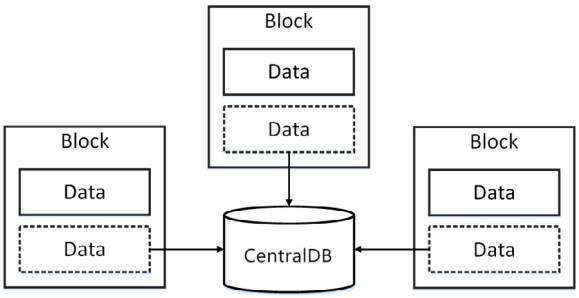

Fig. 1. Separate the data in the blockchain and store it in the central database.

2.1. System structure

Since the blockchain is a distributed ledger which means that every transaction on the blockchain network is recorded on the ledger, so the blockchain data will continue to increase, and at the same time, in order to ensure the data is not tampered with, each node of the blockchain synchronizes the entire network data, resulting in more and more data for a single node, and the queues waiting to confirm transactions are getting longer and longer, making the entire blockchain network bloated [25,26]. Bitcoin founder Nakamoto has set the size of each block to 1 MB in order to reduce the amount of data, but this directly causes the Bitcoin system not being used more widely, because 1 MB is not enough for any organization, especially in the field of big data. The existing mainstream blockchain architecture is subject to its own defects, which makes the blockchain unable to exert its value [27]. At present, there are two ideas for the solution that the block data is too small and the node synchronization data is too much. One is to expand the block and increase the capacity of each block, but as the amount of data in a single block increases, the process of synchronizing data between nodes becomes slower and more bloated. Another method is to reduce the size of the block, although the speed of the node synchronization data can be increased, but this will make the capacity that was not enough to be smaller. These two solutions are like fighting in the left and right hands [28].

Li proposed the etherQL system, which has a separate query layer designed outside the blockchain [29]. The main idea of the system is to copy the blockchain data to an external database, and design the query layer by means of the functional interface provided by the external database. The idea is only to copy the original data of the blockchain to the external database, to improve query efficiency.

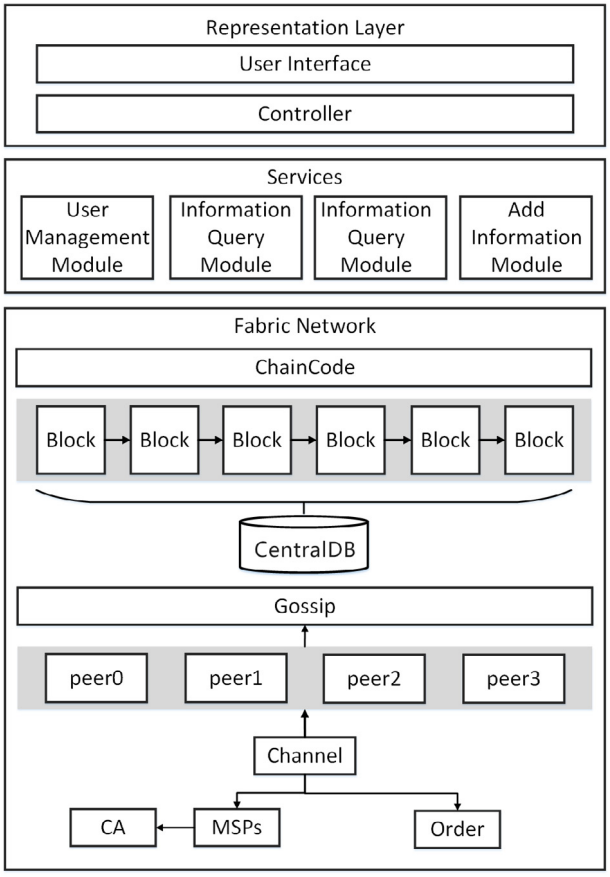

Fig. 2. System structure.

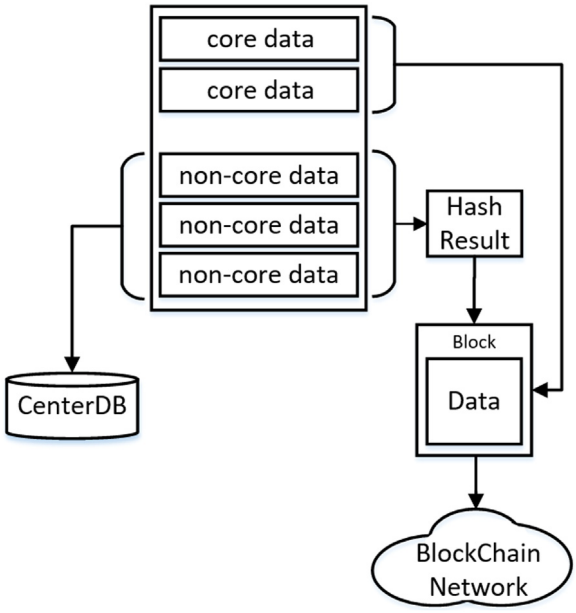

Fig. 3. The process of separating and storing data.

We extend and optimize this idea, and propose a novel on-chain and out-of-chain data model suitable for the Consortium Blockchain. In this idea, as shown in Fig. 1 , some core information is stored in the blockchain, the remaining large amount of unimportant information is stored in a central database outside the blockchain. This method can effectively reduce the amount of data in the blockchain network, and store the hash generated by the out-of-chain data in the block, and then use the Merkle tree to check and trace the data, still can prevent out-of-chain's data being tampered with [30].

Expanding the capacity of the blockchain is a very huge problem in the industry. Many people try to solve it from various aspects, such as sidechains, SegWit, Lightning Network, Hard Fork, etc. The solution basically revolves around several ideas: Keep the existing ceiling unchanged, bypass the restrictions by other means; Directly expand to a certain upper limit such as 2M; And there are some other gradual expansion plans, but these expansion methods are limited to the blockchain framework, because many people believe that since blockchain is a revolution in traditional data storage methods, it should not be related to the traditional data storage method, and it must be completely changed. However, we think that it is necessary to make a compromise and jump out of the blockchain framework to try to find solutions for expanded capacity. So we proposed the idea of combining blockchain and traditional data storage methods.

The system architecture is shown in Fig. 2. The standard MVC software structure is used. From top to bottom is the user view layer, service layer and hyperledger fabric network. The service layer is divided into four functional modules, User Management Module, Information Query Module, Information Modify Module and Add Information Module, chaincode is used to implement intelligent contract operation on blockchain data. CouchDB as a state database to record the results of transaction execution. Most non-core data is stored in a central database, and only core data is stored in the blockchain network. The MSP is a set of encryption mechanisms and protocols for issuing and verifying certificates and identities in the blockchain, it is a pluggable interface. The CA is used to generate certificates and secret keys, initialize the MSP, and the order node acts as a proxy in the network, used to distribute data [31].

2.2. Store data on-chain and out-of-chain

Fig. 3 shows the process of dividing the data and storing into blockchain and central database. Most of the information can be divided into two parts: "core" and "non-core". For example, in the personnel information management system, the core fields in the data structure are "name", "identity number", "the editor of the information", etc. Non-core fields include "date of birth", "graduate college", etc., so when the administrator enters a person information, some fields can be selected as core fields, and these fields will be packaged into one block and stored in the blockchain network, the other unchecked non-core fields perform the SHA256 operation, and the obtained hash results are also stored in the block with the core fields, and all the information is stored in the central database. The data in the blockchain makes the core data tamper-proof and traceable, The non-core data is stored in the central database to effectively reduce the redundancy of the data on the blockchain, and the hash result of the non-core data is stored in the block, when the data is taken out from the central database, the data is performed to SHA256 operation, and the result is compared with the hash result previously stored in the blockchain, if they are the same, it means that the data has not been changed [32].

The SHA256 algorithm can generate a 256-bit long hash value for data of any length [33]. A hash value is a unique and extremely compact numerical representation of a piece of data. If a clear text changes only one letter, the resulting hash value will be completely different, therefore, if the data has been modified, no matter how small changes, the final hash value will be completely different, so the hash value of the data can verify the integrity

of the data [34]. The following briefly describes the SHA256 algorithm process:

(1) Initialization parameters.
Take the first 32 bits of the square root of the first 8 prime numbers (2, 3, 5, 7, 11, 13, 17, 19) in the natural number to get 8 parameters.

$$H_0 = 0x6a09e667;\ H_1 = 0xbb67ae85;$$
$$H_2 = 0x3c6ef372;\ H_3 = 0xa54ff53a;$$
$$H_4 = 0x510e527f;\ H_5 = 0x9b05688c;$$
$$H_6 = 0x1f83d9ab;\ H_7 = 0x5be0cd19;$$

(2) Prepare the message list W_t.

$$W_t = M_t^{(i)}(0 \le t \le 15)$$
$$W_t = \sigma_1^{(256)}(W_{t-2}) + W_{t-7} + \sigma_0^{(256)}(W_{t-15}) + W_{t-16}(16 \le t \le 63)$$

(3) Initialize 8 variables with the intermediate result of each round of hash values.

(4) For $0 \le t \le 63$, execute the compression function.

$$T_1 = h + \sum_{1}^{256}(e) + Ch(e, f, g) + K_t^{256} + W_t$$
$$T_2 = h + \sum_{0}^{256}(a) + M_{aj}(a, b, c)$$
$$h = g; g = f; f = e; e = d + T_1; d = c; c = b;$$
$$b = a; a = T_1 + T_2;$$

(5) Add a compressed block to the current hash value

$$H_0^{(i)} = a + H_0^{(i-1)}, H_1^{(i)} = b + H_1^{(i-1)},$$
$$H_2^{(i)} = c + H_2^{(i-1)}, H_3^{(i)} = d + H_3^{(i-1)},$$
$$H_4^{(i)} = e + H_4^{(i-1)}, H_5^{(i)} = f + H_5^{(i-1)},$$
$$H_6^{(i)} = g + H_6^{(i-1)}, H_7^{(i)} = h + H_7^{(i-1)},$$

Fig. 4 shows the process of querying data. First, the keyword index is used to search the blocks in the blockchain network and the central database, and the data queried from the central database is hashed to obtain the hash result B. The previously saved hash value A is taken from the block in the blockchain network and compared with B. If they are equal, it proves that the data has not been modified.

2.3. Member and organization access mechanisms

Fabric membership is based on a standard X.509 certificate, and the key uses the ECDSA(Elliptic Curve Digital Signature Algorithm), which is a combination of ECC (Elliptic Curve Cryptography) and DSA (Digital Signature Algorithm) [35]. The security of the elliptic curve cryptosystem is based on the intractability of the ECDLP (elliptic curve discrete logarithm problem). Elliptic curve discrete logarithm problem is much more difficult than discrete logarithm problem, the unit bit strength of elliptic curve cryptosystem is much higher than traditional discrete logarithm system. Therefore, in the case of using a shorter key, the ECC can reach the same security level as the DL system. This has the advantage of smaller calculation parameters, shorter keys, faster calculations, and shorter signatures. So elliptic curve cryptography is especially suitable for applications where processing power, memory space, bandwidth, and power consumption are limited. The PKI system is used to issue digital certificates to each

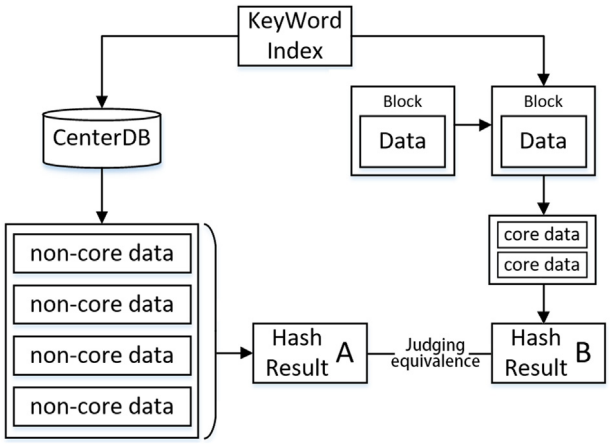

Fig. 4. The process of querying the separated stored data.

member. Only nodes with the same MSP in the channel can use the Gossip protocol for data distribution.

The signature process is as follows:

1. Select an elliptic curve Ep(a,b), and a base point G.;
2. Select the private key k ($k < n$, n is the order of G), using the base point G to calculate the public key K = kG.
3. Generate a random integer r ($r < n$), calculate the point R = rG
4. Taking the original data and the coordinate value x, y of the point R as parameters, and calculating Hash = SHA1 (original data, x, y).
5. Calculate $s \equiv r - Hash * k(mod\ n)$ required in the document;
6. r and s are used as signature values. If one of r and s is 0, re-execute from step 3.

The verification process is as follows:

1. After receiving the message (m) and the signature value (r, s), the receiver performs the following operations.
2. Calculate $sG + H(m)P = (x1, y1), r1 \equiv x1\ mod\ p$
3. Verification equation $r1 \equiv r\ mod\ p$
4. If the equation is true, accept the signature, otherwise the signature is invalid.

Data needs to be shared to release and unearth its value. However, because a lot of data involves the secrets of enterprises and governments, this leads to organizations not sharing their own data with each other.

Because the Consortium Blockchain is pluggable for the organization, allowing multiple certified organizations to join the network to share data. And the tamper-proof security features inherent in blockchains also help reduce the concerns of organizations sharing data, and our proposed on/out-of-chain model still store a large portion of data in local databases. It is acceptable for those traditional companies.

3. Data structure and chain code

According to our user role and function, the data structure of our prototype system is shown in the Fig. 5. When the information is entered, the administrator can check some of the fields as the core field, EditorTime is the timestamp when the current information is stored, and Editor is the administrator who is currently operating the system, these two fields can be used as a proof of the change record of the data. Historys record every change in data and serve as a data traceback.

Each transaction is only valid if it is endorsed according to the endorsement strategy. The endorsement strategy is used to guide the peer how to determine whether the transaction has been approved. When a peer receives a transaction, it invokes the VSCC (Verification System Chaincode) associated with the transaction's Chaincode as part of the transaction validation process to determine the validity of the transaction. A transaction contains endorsement support in one or more peer endorsement nodes. In addition to verifying the endorsement strategy, VSCC also checks if the data version of each Key–Value Pair in the transaction information has changed.

The endorsement strategy has two main components:

1. Principal:P defines the source entity of the expected signature
2. Threshold gate: T has two parameters: integer t (threshold) and n subjects, indicating that t signatures are obtained from these n subjects

for example:

1. T(2, 'A', 'B', 'C') Request signatures of any two endorsement nodes from 'A', 'B', 'C'
2. T(1, 'A', T(2, 'B', 'C')) request a signature from A or from B and C

Fig. 6 is a view of the RBFT structure. In the network, the client sends a request to the node, there is no need to send a message to all nodes, because sending f + 1 is enough. After the node receives the client's request, it will propagate the message so that other nodes know the request message. After each primary node receives the request, it creates a proposal (PRE-PREPARE) and sends it to all other nodes. If other nodes receive the PRE-PREPARE of the primary node, a PREPARE message is returned. Once the node receives the PRE-PREPARE message and 2f PREPARE messages, once the node receives the PRE-PREPARE message and 2f PREPARE messages, the node has enough information to receive the proposal and send a commit message. Once a node receives 2f+1 commit messages, these requests can be sorted and added to the ledger.

Algorithm 1 add personnel information

Input: *Array* Personnel *string* EntityID
Output: bool
if The number of parameters obtained is not 2 **then**
    throw
end if
if The ID number already exists in the database **then**
    throw
end if
applying SHA256 to the data item
json.Unmarshal for Personnel
Store data on block-chain networks
Store data in a central database
Store operation events
    return ;

Algorithm 1 describes how a smart contract (chain code) packs a person's information into a block. First determines the legality of the incoming parameter, and secondly determines whether the person's information has been stored by checking the key field EntityID, If not, perform SHA256 on the data, and insert the result into the data, perform json.Unmarshal processing on the Personnel data structure, and finally save the blockchain and the central database.

Algorithm 2 Query information based on ID number

Input: *string* EntityID *Array* Personnel
Output:
if The number of parameters obtained is not 1 **then**
    throw
end if
Use function GetState to query data and pass in parameter EntityID to assign the result to Personnel
get Hash Result A from Personnel
json.Unmarshal for Personnel
query data from CentralDB and applying SHA256 to the
comparison of Hash Results A and B
use the GetHistoryForKey function to trace historical data and assign the results to iterator
historys ← []*HistoryItem*
hisPer ← *Personnel*
for iterator.HasNext() **do**
    hisData, err := iterator.Next()
    historyItem.TxId ← *hisData.TxId*
    json.Unmarshal for hisData.Value
    historyItem.Personnel ← *hisPer*
    historys ← *append*(*historys*, *historyItem*)
end for
Personnel.Historys ← *historys*
json.Marshal for Personnel
return Personnel;

Algorithm 2 describes how a smart contract (chain code) queries someone's information through the keyword EntityID and tracks the data history. First determine the legality of the incoming parameters, and then use the Getcode function of chaincode to query the data, then perform json.Unmarshal processing, Obtain the result A previously stored in the block, Obtain the hash result A stored in the previous database, and then take the data from the central database, and perform SHA256 operation to obtain the hash result B, and determine whether the hash results A and B are equal. If they are equal, It means that the central database data has not been modified, then merge the data from blockchain and central database. Use the chaincode GetHistoryForKey function to query the traceability data and store it in the iterator, then loop through the iterator to store the data in the historys array.

4. Experimental results and analysis

We developed a prototype system based on the hyperledger fabric to verify that our proposed idea of separating and storing data is indeed effectively applicable to such big data management systems as personnel information management. The system runs on a Ubuntu 18.04 (64-bit) virtual machine, Intel(R) Core(TM) it-470HQ CPU @ 2.50 GHz processor and 8 GB RAM, and uses the MySQL Ver 14.14 simulation central database.

Fig. 7 is a demonstration of the operation in the prototype system, showing the results of querying a person's information based on the ID number, the top table shows the change of personnel information, and each column marks the operator of the change and the operation time, which realizes the traceability of the data.

Next, observe the changes in the read and store response time after separately storing the data and not separating it. The size of the space occupied by a single person is 50 kB, In the separate storage scheme, 40 kB of data is stored in the central database, 10 kB is stored in the blockchain network, and in the non-separate storage scheme, all 50 kB of data is stored in the blockchain network. Starting from the data volume of 2000 people, the number

```
type Personnel struct{
    ObjectType      string
    Name            string
    Gender          string
    Nation          string
    EntityID        string
    Place           string
    BirthDay        string
    EntryDate       string
    EditTime        string
    GraduateSchool  string
    Major           string
    EducationDegree string
    Position        string
    PositionLevel   string
    Department      string
    WorkNumber      string
    Editor          string
    Historys        []HistoryItem
}
```

Fig. 5. Data structure.

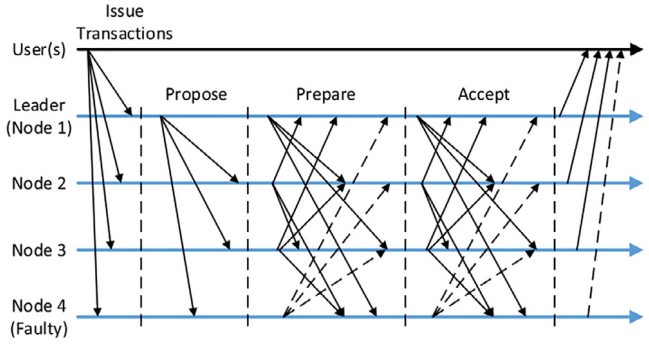

Fig. 6. RBFT structure.

name	gender	nation	ID card	place	birthDay	entryDate	graduateSchool
neptune	male	han	0001	QingDao	1994-10-9	2018-3-8	QingDao University
neptune	male	han	0001	QingDao	1994-10-9	2018-3-8	QingDao University
neptune	male	han	0001	QingDao	1994-10-9	2018-3-8	QingDao University

major	educationDegree	position	positionLevel	department
accounting	master	software engineer	3	Technology
accounting	master	Senior Software Engineer	5	Technology
accounting	master	Senior Software Engineer	5	Technology

workNumber	editTime	editor
1111	2019-03-15 20:46:03	admin1
1111	2019-03-15 23:29:32	admin1
2222	2019-03-15 23:30:49	admin2

Name：	neptune	GraduateSchool：	QingDao University
Gender：	male	Major：	accounting
Nation：	han	EducationDegree：	master
ID Card：	0001	Position：	Senior Software Engineer
Place：	QingDao	PositionLevel：	5
BirthDay：	1994-10-9	Department：	Technology
EntryDate：	2018-3-8	WorkNumber：	2222
EditTime：	2019-03-15 23:30:49	Editor：	admin1

Modify information Back to home page

Fig. 7. Data result.

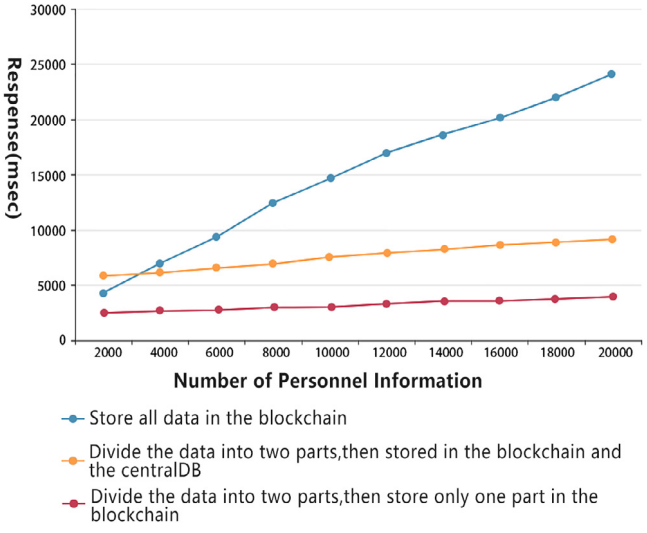

Fig. 8. The relationship between the storage data response time and the amount of data on the chain. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

is incremented by 2000 people until 20,000. The blue curve in the figure is to store all the data in the blockchain network. The red curve is to store only 10 kB of data into the blockchain, in order to observe the impact on performance only by reducing the size of the stored data. The yellow curve stores 10 kB of data into the blockchain, and the remaining 40 kB of data is stored in the central database.

Fig. 8 shows the response time for add data as the amount of data on the blockchain increases. It can be seen that reducing the amount of data can effectively reduce the storage response time. Because the separate data storage needs to store the data in the blockchain and the local database respectively, there is one more step than storing all the data in the blockchain network, therefore, when the amount of data is small, the response time of the separate storage will be higher than not separating storage, but storing data in a local database does not require consensus verification and node data synchronization. As the amount of data increases, the advantages of separate storage gradually emerge.

Fig. 9 shows the change in query response time as the amount of data increases. As with storage, reducing the amount of data can be very effective in reducing query time. When using the on-chain and out-of-chain model, there are two steps in a query, including querying the blockchain network and the central database, because there is one more step than the traditional query method, so when the amount of data is small, Compared with the traditional method, the on-chain and out-of-chain model not have the advantage, but as the amount of data increases, the advantages are more obvious.

Fig. 10 is a performance analysis and comparison of the two methods of storing all data in blockchain and storing data us

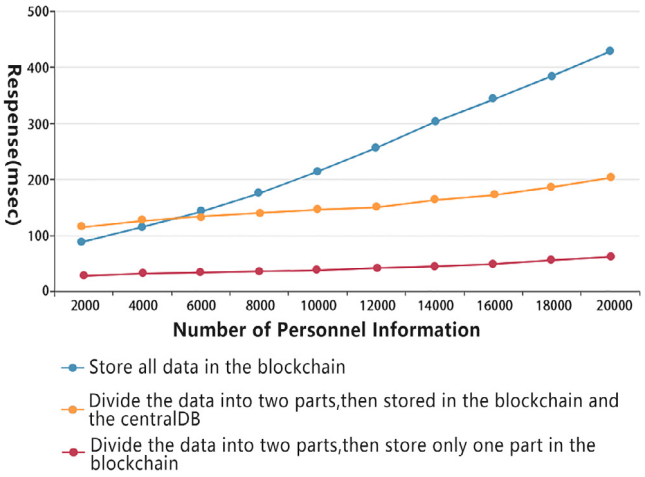

Fig. 9. The relationship between the query data response time and the amount of data on the chain.

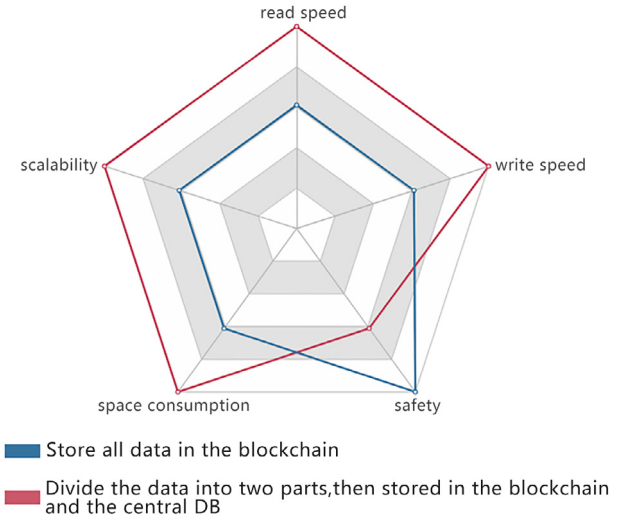

Fig. 10. Performance comparison of the two schemes.

on/out-of-chain model. It can be analyzed from Figs. 8 and 9 that in the application scenario such as personnel data management, which needs to store a large amount of data, the write and read performance of the separate data storage is better than the non-separated data storage, especially with the more data, the advantage is more obvious. Since the data is stored separately, the amount of data on the blockchain network can be actually reduced, and the central database is stored locally, so it is not limited by the storage capacity, and can be expanded at any time, Therefore, the separate storage scheme reduces the amount of data on the blockchain network, which has many advantages, such as speeding up the writing and query speed, improving system scalability and space utilization, etc., Because the core idea of blockchain is to store data in a distributed manner to ensure data security, and our model combines the blockchain with the traditional data storage scheme, but because some of the data is stored in a central database, the security is lower than the scheme of storing all the data in the blockchain. This method is a compromise between new and old technologies, it is only less secure than storing data in a blockchain network, but it is still more secure than traditional data storage solutions. Because the most important part of the data is still stored on the blockchain network, while other data is stored in the local database, according to the method we designed, the data will get a hash value and stored in the blockchain, when you need to query the data, it will compare whether the two hash values change, so even though it is stored in the local database, the security is still higher overall. Therefore, the final result must be better than the old technology, that is, better than the technology that is now widely used, but to a certain extent, it cannot fully achieve the effect of the ideal new technology.

5. Summary and outlook

In this paper, in order to solve the security problem of personnel information management in big data, we propose a framework of personnel information management system based on blockchain, and develop a prototype system to verify the feasibility of this model, and propose a novel solution for separating and storing data to solve the problem of blockchain information redundancy and insufficient storage space, at the same time have the advantage of traditional database that can store large amounts of data. This model can effectively solve problems such as information leakage and tampering. Because the organization of the Consortium Blockchain is pluggable, organizations can apply to join the system at any time to achieve data sharing. This idea of separating and storing data is scalable, we believe that almost all information management fields can refer to this idea. In this article, we only use personnel information management as a breakthrough, we will study this method in depth and try to extend it to more fields.

At present, the blockchain technology is still in the early stage of verification, and there is still no large-scale application scenario, especially in the big data field, the blockchain is not widely used because it is limited by its own bottleneck such as limited data capacity. However, it is foreseeable that the non-tampering and traceability of the blockchain naturally have the application advantages of industries such as finance and credit reporting. The Private Blockchain and the Consortium Blockchain have relatively strict access mechanisms and regulatory measures, and can be applied to data storage mode proposed by this paper, which better solves the problem of large-scale data storage and will become the main blockchain technology in the future.

Acknowledgement

This work was supported in part by the Shandong Provincial Natural Science Foundation (ZR2017QF015)

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Z.A. Soomro, M.H. Shah, J. Ahmed, Information security management needs more holistic approach: A literature review, Int. J. Inf. Manage. 36 (2) (2016) 215–225.
[2] W. Cui, N. Zhang, Research and development of filing management system of school personnel information based on web, J. Appl. Sci. Eng. Innov. 4 (4) (2017) 127–130.
[3] R. Nisar, S.G. Sahar, Security and privacy issues, in: 2019 2nd International Conference on Computing, Mathematics and Engineering Technologies (ICoMET), IEEE, 2019, pp. 1–6.
[4] L. Cheng, F. Liu, D.D. Yao, Enterprise data breach: causes, challenges, prevention, and future directions, Wiley Interdiscip. Rev.: Data Min. Knowl. Discov. 7 (5) (2017).
[5] M. Sokolova, S. Matwin, Personal privacy protection in time of big data, in: Challenges in Computational Statistics and Data Mining, Springer, 2016, pp. 365–380.

[6] M. Wen, S. Yu, J. Li, H. Li, K. Lu, Big data storage security, in: Big Data Concepts, Theories, and Applications, Springer, 2016, pp. 237–255.
[7] T.T.A. Dinh, R. Liu, M. Zhang, G. Chen, B.C. Ooi, J. Wang, Untangling blockchain: A data processing view of blockchain systems, IEEE Trans. Knowl. Data Eng. 30 (7) (2018) 1366–1385.
[8] Z. Zheng, S. Xie, H. Dai, X. Chen, H. Wang, An overview of blockchain technology: Architecture, consensus, and future trends, in: 2017 IEEE International Congress on Big Data (BigData Congress), IEEE, 2017, pp. 557–564.
[9] D. Puthal, N. Malik, S.P. Mohanty, E. Kougianos, C. Yang, The blockchain as a decentralized security framework [future directions], IEEE Consum. Electron. Mag. 7 (2) (2018) 18–21.
[10] G. Karame, S. Capkun, Blockchain security and privacy, IEEE Secur. Priv. 16 (4) (2018) 11–12.
[11] P. Dunphy, F.A. Petitcolas, A first look at identity management schemes on the blockchain, IEEE Secur. Priv. 16 (4) (2018) 20–29.
[12] R. Beck, M. Avital, M. Rossi, J.B. Thatcher, Blockchain Technology in Business and Information Systems Research, Springer, 2017.
[13] X. Wang, L. Feng, H. Zhang, C. Lyu, L. Wang, Y. You, Human resource information management model based on blockchain technology, in: 2017 IEEE Symposium on Service-Oriented System Engineering (SOSE), IEEE, 2017, pp. 168–173.
[14] J.I. Zahid, A. Ferworn, F. Hussain, Blockchain: A technical overview, IEEE Internet Policy Newsl. (2018) 1–3.
[15] S. Nakamoto, et al., Bitcoin: A Peer-To-Peer Electronic Cash System, 2008.
[16] T.T.A. Dinh, J. Wang, G. Chen, R. Liu, B.C. Ooi, K.-L. Tan, Blockbench: A framework for analyzing private blockchains, in: Proceedings of the 2017 ACM International Conference on Management of Data, ACM, 2017, pp. 1085–1100.
[17] K. Lei, Q. Zhang, L. Xu, Z. Qi, Reputation-based byzantine fault-tolerance for consortium blockchain, in: 2018 IEEE 24th International Conference on Parallel and Distributed Systems (ICPADS), IEEE, 2018, pp. 604–611.
[18] E. Karafiloski, A. Mishev, Blockchain solutions for big data challenges: A literature review, in: IEEE EUROCON 2017-17th International Conference on Smart Technologies, IEEE, 2017, pp. 763–768.
[19] Z. Li, J. Kang, R. Yu, D. Ye, Q. Deng, Y. Zhang, Consortium blockchain for secure energy trading in industrial internet of things, IEEE Trans. Ind. Inform. 14 (8) (2017) 3690–3700.
[20] E. Androulaki, A. Barger, V. Bortnikov, C. Cachin, K. Christidis, A. De Caro, D. Enyeart, C. Ferris, G. Laventman, Y. Manevich, et al., Hyperledger fabric: a distributed operating system for permissioned blockchains, in: Proceedings of the Thirteenth EuroSys Conference, ACM, 2018, p. 30.
[21] C. Cachin, Architecture of the hyperledger blockchain fabric, in: Workshop on Distributed Cryptocurrencies and Consensus Ledgers, Vol. 310, 2016.
[22] P. Thakkar, S. Nathan, B. Viswanathan, Performance benchmarking and optimizing hyperledger fabric blockchain platform, in: 2018 IEEE 26th International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS), IEEE, 2018, pp. 264–276.
[23] A. Baliga, N. Solanki, S. Verekar, A. Pednekar, P. Kamat, S. Chatterjee, Performance characterization of hyperledger fabric, in: 2018 Crypto Valley Conference on Blockchain Technology (CVCBT), IEEE, 2018, pp. 65–74.
[24] R. Shi, Y. Gan, Y. Wang, Evaluating scalability bottlenecks by workload extrapolation, in: 2018 IEEE 26th International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS), IEEE, 2018, pp. 333–347.
[25] T. Ahram, A. Sargolzaei, S. Sargolzaei, J. Daniels, B. Amaba, Blockchain technology innovations, in: 2017 IEEE Technology & Engineering Management Conference (TEMSCON), IEEE, 2017, pp. 137–141.
[26] L. Yue, H. Junqin, Q. Shengzhi, W. Ruijin, Big data model of security sharing based on blockchain, in: 2017 3rd International Conference on Big Data Computing and Communications (BIGCOM), IEEE, 2017, pp. 117–121.
[27] T. McConaghy, R. Marques, A. Müller, D. De Jonghe, T. McConaghy, G. McMullen, R. Henderson, S. Bellemare, A. Granzotto, Bigchaindb: a scalable blockchain database. white paper, BigChainDB, 2016.
[28] J. Göbel, A.E. Krzesinski, Increased block size and bitcoin blockchain dynamics, in: 2017 27th International Telecommunication Networks and Applications Conference (ITNAC), IEEE, 2017, pp. 1–6.
[29] Y. Li, K. Zheng, Y. Yan, Q. Liu, X. Zhou, Etherql: a query layer for blockchain system, in: International Conference on Database Systems for Advanced Applications, Springer, 2017, pp. 556–567.
[30] R. Shi, Y. Wang, Cheap and available state machine replication, in: 2016 USENIX Annual Technical Conference (USENIX ATC 16), 2016, pp. 265–279.
[31] Q. Nasir, I.A. Qasse, M. Abu Talib, A.B. Nassif, Performance analysis of hyperledger fabric platforms, Secur. Commun. Netw. 2018 (2018).
[32] M. Di Pierro, What is the blockchain?, Comput. Sci. Eng. 19 (5) (2017) 92–95.
[33] D. Rachmawati, J. Tarigan, A. Ginting, A comparative study of message digest 5 (md5) and sha256 algorithm, in: Journal of Physics: Conference Series, Vol. 978, IOP Publishing, 2018, p. 012116.
[34] B. Applebaum, N. Haramaty-Krasne, Y. Ishai, E. Kushilevitz, V. Vaikuntanathan, Low-complexity cryptographic hash functions, in: 8th Innovations in Theoretical Computer Science Conference (ITCS 2017), Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017.
[35] Y. Lindell, Fast secure two-party ecdsa signing, in: Annual International Cryptology Conference, Springer, 2017, pp. 613–644.

Jian Chen is currently a graduate student in the School of Data Science and Software Engineering at Qingdao University. His research interests include blockchain and virtual reality. In 2017, he obtained a bachelor's degree from Xinjiang University. In 2018, he went to the University of Delaware for a semester of study. In the same year, he won the second prize of the National Computer Contest in China. He has rich experience in web and mobile software development and helped the Chinese Embassy to develop software.

Zhihan Lv is currently a Associate Professor of Qingdao University, China. He was Research Associate at University College London (UCL). He has been an assistant professor in Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences since 2012. He received his PhD from Paris7 University and Ocean University of China in 2012. He worked in CNRS (France) as Research Engineer, Umea University/ KTH Royal Institute of Technology (Sweden) as Postdoc Research Fellow, Fundacion FIVAN (Spain) as Experienced Researcher. He was a Marie Curie Fellow in European Union's Seventh Framework Program LANPERCEPT. His research mainly focuses on Multimedia, Augmented Reality, Virtual Reality, Computer Vision, 3D Visualization & Graphics, Serious Game, HCI, Big data, and GIS. He has contributed 200+ papers in the related fields on journals such as Plos one, ACM TOMM, and conference such as ACM MM, ACM CHI, ACM Siggraph, ICCV, IEEE Virtual Reality.

He is the guest editors for IEEE Transactions on Industrial Informatics, Future Generation Computer Systems, Neurocomputing, Neural Computing and Applications, Computers & Electrical Engineering, IET Image Processing, Multimedia Tools and Applications, Journal of Intelligent and Fuzzy Systems. He is Program Committee member of ACM IUI 2015 & 2016 &2019, IEEE BIGDATA4HEALTH Workshop 2016, IEEE/CIC WIN Workshop 2016, IIKI2016, WASA2016,IEEE PDGC2016, and ACM SAC2017-WCN Track. He has also served as a reviewer for journals such as IEEE Transaction on Multimedia, IEEE Multimedia, Neurocomputing, Elsevier Computer Networks, Springer Telecommunication Systems, Multimedia Tools and Applications, KSII Transactions on Internet and Information Systems, Journal of Medical Internet Research, Intelligent Service Robotics, PRESENCE: Tele-operators and Virtual Environments, and conference.

Houbing Song (M'12-SM'14) received the Ph.D. degree in electrical engineering from the University of Virginia, Charlottesville, VA, in August 2012, and the M.S. degree in civil engineering from the University of Texas, El Paso, TX, in December 2006. In August 2017, he joined the Department of Electrical, Computer, Software, and Systems Engineering, Embry-Riddle AeronauticalUniversity, Daytona Beach, FL, where he is currently an Assistant Professor and the Director of the Security andOptimization forNetworkedGlobe Laboratory (SONG Lab, www.SONGLab.us). He served on the faculty of West Virginia University from August 2012 to August 2017. In 2007 he was an Engineering Research Associate with the Texas A&M Transportation Institute. He serves as an Associate Technical Editor for IEEE Communications Magazine. He is the editor of six books, including Big Data Analytics for Cyber-Physical Systems: Machine Learning for the Internet of Things, Elsevier, 2019, Smart Cities: Foundations, Principles and Applications, Hoboken, NJ: Wiley, 2017, Security and Privacy in Cyber-Physical Systems: Foundations, Principles and Applications, Chichester, UK: Wiley-IEEE Press, 2017, Cyber-Physical Systems: Foundations, Principles and Applications, Boston, MA: Academic Press, 2016, and Industrial Internet of Things: Cybermanufacturing Systems, Cham, Switzerland: Springer, 2016. He is the author of more than 100 articles. His research interests include cyber-physical systems, cybersecurity and privacy, internet of things, edge computing, big data analytics, unmanned aircraft systems, connected vehicle, smart and connected health, and wireless communications and networking.