2022 21st International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES)

Security Protection of Research Sensitive Data Based on Blockchain

Cheng Cheng, Zixiang Liu, Feng Zhao
China Ship Scientific Research Center
TAIHU Laboratory of Deepsea Technological Science
Wuxi, China
e-mail: chengc@cssrc.com.cn

Xiang Wang, Feng Wu
Hangzhou Fengshun Technology Information Service Co. Ltd
Hangzhou, China
e-mail: wangxiang@hzfengshun.com

***Abstract*—In order to meet the needs of intellectual property protection and controlled sharing of scientific research sensitive data, a mechanism is proposed for security protection throughout "transfer, store and use" process of sensitive data which based on blockchain. This blockchain bottom layer security is reinforced. First, the encryption algorithm used is replaced by the national secret algorithm and the smart contract is encapsulated as API at the gateway level. Signature validation is performed when the API is used to prevent illegal access. Then the whole process of data up-chain, storage and down-chain is encrypted, and a mechanism of data structure query and data query condition construction based on blockchain smart is provided to ensure that the data is "usable and invisible". Finally, data access control is ensured through role-based and hierarchical protection, and the blockchain base developed has good extensibility, which can meet the requirement of sensitive data security protection in scientific research filed and has broad application prospects.**

***Keywords-** sensitive data; controlled sharing; data encryption;access control*

I. INTRODUCTION

There are usually multiple technical status in the designation and production of complex equipment, and the whole process is managed by multiple departments or institutes. Due to the management system, confidential rules, level of informatization and intellectual property protection, the data exchange for technical status transformation is based on the offline mode for the most time. Numerous data is kept by specific institutes which can hardly be shared sufficiently. When changes happen with the existing data, other cooperating institutes cannot be informed in time. What's more, the interaction may be incorrect, without clear responsibility, as shown in Fig1. These problems lead to the lack of whole period data of equipment R&D, and the value of data cannot be fully used.

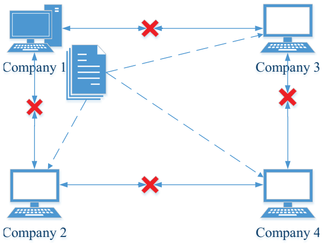

Figure 1. Current situation of scientific research data interaction

The technical features of blockchain include high reliability, privacy preserving, tamperproof, action traceability and self-establishment of trust. These features make blockchain applicable to the authentication of digital properties, and it can also be applied to the controlled sharing and application of data with smart contracts and encryption-decryption. There are already a vast of research have been carried out on the application of blockchain in data transactions. Banerjee P[1] discussed the problems in data transaction security and introduced blockchain to solve it. Wang, Zhang and Li[2-5] launched research on the combination of data transaction and blockchain to improve the security of data and the transparency of transaction. As an open source technology from abroad, many core functions are based on foreign algorithms [6-9], including Hyperledger Fabric, which is the most popular blockchain framework at present. Though it is available for directly application in public zone, R&D of equipment has much higher demand on the controlling of data. Cao[10] studied on replacing the original encryption algorithms with domestic SM series algorithm.

The introduction on implements of blockchain to equipment R&D remains blank till now. The protection method of data in transmission, storage and application based on blockchain is introduced in this paper, which could overcome the shortcomings of classical data management while meeting the demand of data security in equipment R&D.

II. OVERALL DESIGNATION

To ensure the security of data in sharing and application, the security system shall be established on basement layer, storage layer, transmission layer and application layer, as shown in Fig2.

The overall design consists of four parts.

1) Basement layer: replace the encryption algorithms of original Fabric, extend the consensus algorithms and retrofit the peer management functions to enhance the security of blockchain framework.

2) Storage layer: ensure the security of data with business channels, distributed content addressing and privacy zone technics.

3) Transmission layer: secure the access and transmission of data through certification, data service, auto encryption-decryption and certification verification API with key hosting on blockchain.

4) Application layer: enable the authentication of data, authorization of access and the set of request rules, the male data appliable while invisible and secure the data property.

2473-3636/22/$31.00 ©2022 IEEE
DOI 10.1109/DCABES57229.2022.00073

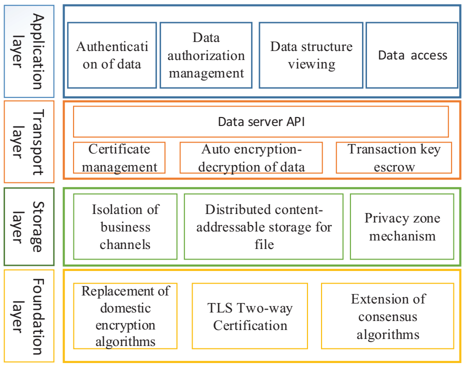
Application layer
Authentication of data
Data authorization management
Data structure viewing
Data access
Transport layer
Data server API
Certificate management
Auto encryption-decryption of data
Transaction key escrow
Storage layer
Isolation of business channels
Distributed content-addressable storage for file
Privacy zone mechanism
Foundation layer
Replacement of domestic encryption algorithms
TLS Two-way Certification
Extension of consensus algorithms

Figure 2. Overall design of data security protection

III. Security of Basement Layer

A. *Replacement of encryption algorithms*

The international standard encryption algorithms are applied as default algorithms of Fabric. To improve the security of blockchain, it is important to replace the default algorithms with the domestic SM series. The correspondence between default encryption algorithms in fabric and SM series is shown in table 1.

TABLE I. Correspondence between default encryption algorithms in fabric and SM series

Purpose	**Default**	**SM Series**
Hash calculation	SHA-2	SM3
User ID	SHA-2	SM3
Signature/verification	ECDSA	SM2
CA certification	ECDSA	SM2
Data transmission	RSA	SM2
Data storage	AES or RSA	SM4 or SM2

B. *Extension of consensus algorithms*

The consensus algorithms of Fabric are extended in this paper, with three types available, including Practical Byzantine Fault Tolerance (PBFT), Fast Byzantine Fault Tolerance (FBFT) and Raft. These algorithms are able to be switched for specific scenarios or different demand for security and performance.

1) FBFT

The PBFT is originally applied in Fabric with $3f+1$ orders, in which f stands for the maximum number of tolerated error orders. Up to 1/3 error orders could be tolerated with PBFT, which is appliable for untrusted multi-party transactions.

Though widely applied, there are drawbacks in PBFT. To overcome the error or cheating of the leader peer, PBFT applied full point-to-point communication to monitor various abnormal behaviors. This has led to the high complexity of communication which reached $O(n^2)$ and a lot of additional signature verification, causing heavy burden to the system and reducing of effect. Moreover, the PBFT is unavailable for service during the election of the leader peer. If the new leader peer cheats or failures, continuous election may happen, which could lead to the blockchain service unavailable.

In order to solve these questions, the FBFT consensus algorithm is developed. The drawbacks of PBFT are sufficiently fixed through optimizing the process of consensus, and the stability of service is ensured during the election of leader peer or error happens. The complexity of communication is reduced from $O(n^2)$ to $O(n)$ by cancelling unnecessary signature verifications, the process of consensus is simplified and the efficiency is improved.

2) Raft

Raft is a type of Crash Fault Tolerance (CFT), which can tolerate half of the orders with error. When $(N+1)/2$ (round up) of N order are proper functioning, the process of consensus is guaranteed to run correctly.

The two consensus algorithms apply to different application scenarios.

- FBFT

The FBFT needs 4-10 orders, with up to $(N\text{-}1)/3$ (N is he total number of orders) error orders could be tolerated during the process of ordering. When the demand for efficiency is high, the FBFT is suitable.

- Raft

Up to $(N+1)/2$ (N is he total number of orders) of N error orders could be tolerated during the process of ordering. When the demand for reliability is high, the Raft is suitable.

IV. Security of Storage

A. *Isolation of business channels*

Multiple institutes in the industry can form an alliance in the blockchain. Under the alliance, several different institutes can establish different business channels according to different business categories. Each channel has an independent ledger, and only channel members can share it. The channels can ensure the formation of a specific private network between members, on which transactions are performed in a confidential manner, and isolated from external unrelated individuals.

B. *Privacy zone mechanism*

In order to solve the problem of privacy data, privacy zone is provided in the blockchain framework. Privacy zone has the functions including encryption, signature, identity authentication and data exclusive zone. The privacy zone is only created on the peers within the data provider organization, and other alliance members do not have any source data backup. The data provider can dynamically set the permission of the privacy zone, and open the privacy area data to the designated consortium chain members.

C. *Distributed content-addressable storage for file*

Distributed content-addressable storage enables the fast storage of file data on the blockchain. The storage space of the blockchain is divided into fixed-size storage units, each unit has its own storage category, access authority, area and other attributes, and each storage unit has an independent access address on the blockchain.

Distributed content addressing storage assigns a unique hash to each file, through which addressing based on file content can be achieved. The blockchain uses the hash as an index, and stores the file slices in a series of storage units in series according to certain rules, and establishes version management for each file. When querying a file, the search service is based on the hash of the file.

V. Security of Transmission

A. Certificate management

1) Certificate request

After completing the client registration in the application system, the blockchain membership certificate is applied by calling the API interface of the blockchain. When applying for a member certificate, the blockchain service will verify the organization of members. According to the predetermined MSP management strategy, a unique certificate will be generated on the organization CA and bound with the user information to confirm the registration, and the certificate will be returned to the application system and saved.

2) Certificate application

When the business system sends a transaction request to the blockchain API, it needs to use the blockchain certificate to sign the transaction request message. After receiving the transaction request, the API verifies the signature in the message. Only when the transaction passes the signature verification, the gateway will perform subsequent transaction processing on the transaction request message. The SHA256WithECDSA encryption algorithm is applied to sign the total content of communication, which can effectively avoid replay attacks and command counterfeiting attacks.

B. Transaction key escrow

After the new client completes the registration and obtains the certificate, the user's transaction key is escrowed by calling the API provided by the blockchain. Under the public key escrow mode, the data submitted by the user will be automatically encrypted by the blockchain and uploaded to the blockchain. When requesting data from the blockchain, the smart contract performs authority verification according to the identity of the requester that has passed the authentication. After the verification is passed, the public key of the data provider is obtained and the data is automatically decrypted and sent to the blockchain for TLS communication encryption and return.

C. Auto encryption-decryption of data

In order to sufficiently cope with the different encryption requirements of various types of data, a unified data encryption identification specification has been formulated. The data is identified according to the specifications. Once the data is identified, the blockchain will automatically encrypt it with the user's private key during the data storage process, as shown in Fig.3. When the data is downloaded, after the blockchain obtains the certificate data through authority authentication, the data is automatically decrypted with the escrowed public key. The decrypted data is encrypted by the blockchain TLS and then transmitted to the gateway node. The gateway node decrypts it again and sends it to the application system, as shown in Fig.4. Because the blockchain gateway node and the application system are deployed in the same network domain, the data outside the domain can be guaranteed.

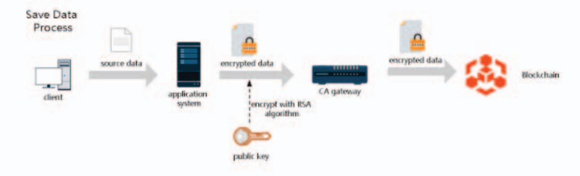

Figure 3. Automatic encryption process of data storage to blockchain

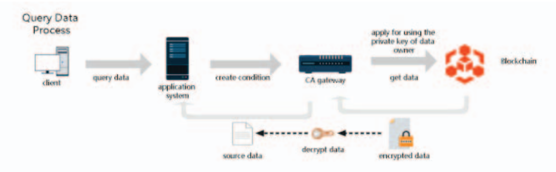

Figure 4. Automatic decryption process of data obtained from blockchain

VI. Security of Application

A. Authentication of data

Through the implementation of data self-registration, digital DNA certificate issuance, and data application tracking on the blockchain, the complete data authentication management is enabled.

1) Self-registration of data

The blockchain provides non-heterogeneous rest SDK and convenient algorithm tools to enable the self-registration of data. On the premise that the original data does not need to be uploaded to the blockchain, the basic information and the hash of the core file are used to perform the self-determination and registration of data on the blockchain. The core file hash has the technical features of irreversible, leak-free and unique.

2) Digital DNA certificate

Through the submitted core information, including the provider's name, registration time and name, the core file abstract hash is coordinated, and a unique digital DNA certificate is generated by encryption and packaging according to the standard protocol rules. The certificate carries the complete and true information of the data, and has the security ability to verify the authenticity of the data, and will be used as the unique identification for copyright authorization in the future.

3) Data application tracking

After the data is authenticated on the blockchain, when it is shared and applied, the blockchain will record the whole process of its application, including users ID, time, details and usage evaluation. Tracking the whole process of data application can avoid data infringement and protect the rights and interests of data owners. Based on a large number of application records and evaluations, it is possible to analyze and evaluate the pros and cons of data to help optimize the value of data.

B. Data authorization management

The data authorization management service supports setting read and write limits for data. The data access permission attributes that can be set by the service include authorized organizations, authorized departments, authorized users, authorization expiration time, and data usage times.

C. Data structure viewing and querying

In order to ensure that the data stored on the blockchain can be effectively applied without the original data being obtained, data structure viewing and querying API is introduced in this paper by imitating the data query principle of relational databases, and users could apply data through programs.

The program is released in the application system after being reviewed, and is provided to users online, making the data appliable while invisible.

The data structure viewing function on the blockchain is similar to the database structure viewing function of the relational database. The detailed composition of the data on the chain can be viewed, thereby laying the foundation for data analysis and data usage. The data query condition construction API provides functions similar to SQL statements in relational databases, allowing users to construct query conditions for query fields and obtain suitable available data.

VII. Demonstration and Analysis

To test and verify the data sharing and application method introduced in this paper, a demonstration has been established, including a blockchain framework and application system. An array of ship resistance data is applied for the system function test, and the ship resistance prediction program based on dynamic surrogate modelling, which needs external data to execute, is used as a demo.

Multiple users are created on peer servers as simulation of different cooperation partners for the test. The demo data is assigned to different simulated partners and upload to the blockchain and stored with authentication certification after duplication checking. All users have the authority to see the data structure on the blockchain as shown in Fig.5, but not the contents.

Figure 5. Data structure view function

The users of data can request and apply data according to the data structure and usage requirements. In this demonstration,authorized data is obtained from the blockchain and applied by the program to build surrogate models for prediction.

The usage log of data will be saved on the blockchain when the prediction in done. The data provider could check the usage log as shown in Fig.6, including the identity of users, time and program applied.

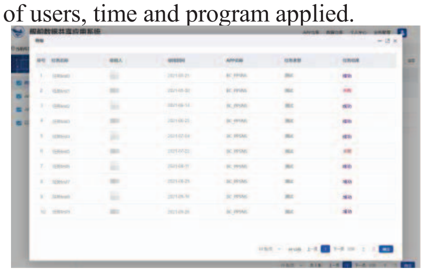

Figure 6. View data usage records

To verify the security and efficiency of blockchain framework, run time of the consensus algorithms and SM algorithms is recorded and analyzed. Configuration of the blockchain server used for the verification was 32 cores CPU and 64GB RAM wit Linux OS. The Fabric network has two organizations, four peers and five orders.

A. Comparison of encryption algorithms

User information strings of 32/64/128/256 bytes is applied to the generation of signatures with SM2, ECDSA and RSA. The operation was repeated for 1000 times to analyze the average time of different algorithms. The SM2 has the highest efficiency, with ECDSA very close to it, and the RSA was the lowest, as shown in Fig.7.

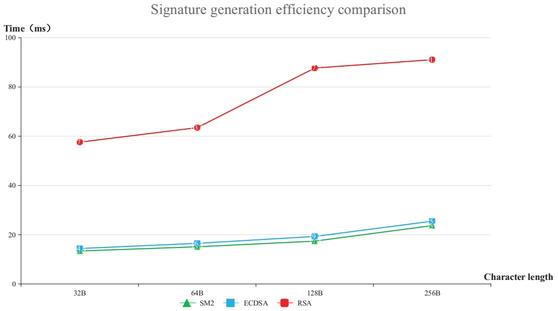

Figure 7. Comparison of signature generation efficiency

Basic data information of 32/64/128/256 bytes is applied to the generation of summary hash with SM3 and SHA-256. The operation was repeated for 1000 times to analyze the average time of different algorithms, and the result shows that the efficiency of SM3 is lower than SHA-256, as shown in Fig.8.

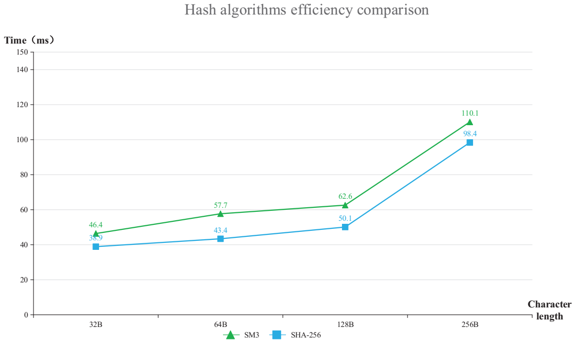

Figure 8. Comparison of hash algorithm efficiency

B. Comparison of consensus algorithms

The three most frequently used cases in the application system including save Data (data storage), get DataFileList (obtain data from blockchain) and auth (data authentication) were applied through caliper-benchmarks component to test the performance of consensus algorithms. All the three cases were applied as concurrent requests for 1000 times, and the result shows that the FBFT has the highest efficiency, as shown in Fig.9.

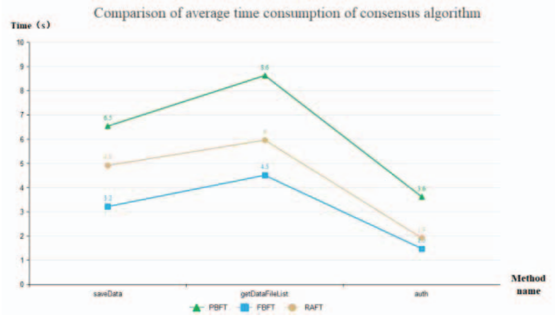

Figure 9. Comparison of average time consumption of consensus algorithm

VIII. Conclusion

It is verified by the demonstration that the blockchain based R&D data protection method introduced in this paper has high efficiency and reliability. This method is able to sufficiently ensure the security of data during transmission, storage and application, and secure the user information and usage log as well. It is a common technology for the authentication and controlled sharing of data, which can be widely used in all scenarios of digital property transaction and sufficiently boost the establishment of the new R&D mode based on cooperation and sharing.

References

[1] BANERJEE P, RUJ S. Blockchain enabled data marketplace - design and challenges[J]. arXiv preprint, 2018, arXiv:1811.11462.

[2] ZHANG Z W, WANG G R, XU J L, et al. Survey on data management in blockchain sys-tems[J]. Journal of Software, 2020, 31(9): 2903-2925.

[3] WANG J W, ZHENG Z Z, WU F, et al. Blockchain based data marketplace[J]. Big Data Research, 2020, 6(3):21-35.

[4] ZHANG Zhao, TIAN J X, JIN C Q. On-chain witness and off-chain transmission trustworthy data sharing platform[J]. Big Data Research, 2020,6(5):106-117.

[5] LI Yuan, GAO N, SUN J, ZHAO H Q. Research and exploration of big data transaction model based on blockchain[J]. Big Data Research, 2021,7(4):37-48.

[6] Hyperledger-fabricdocs documenta-tion[EB].

[7] Hyperledger-fabric-ca documentation[EB].

[8] Hyperledger Fabric SDKs[EB].

[9] ZHANG Q H. Research on identification and access control in blockchain[D]. Beijing: Beijing Jiaotong University, 2018.

[10] CAO Qi, RUAN S H, CHEN X S, et al. Em-bedding of national cryptographic algorithm in hyperledger fabric[J]. Chinese Journal of Net-work and Information Security, 2021,7(1):65-75