Big Data Model of Security Sharing Based on Blockchain

LI Yue
University of Electronic Science and Technology of China
Chengdu, China
ly_uestc@outlook.com

HUANG Junqin
University of Electronic Science and Technology of China
Chengdu, China
huangjunqin@std.uestc.edu.cn

QIN Shengzhi
University of Electronic Science and Technology of China
Chengdu, China

WANG Ruijin*
University of Electronic Science and Technology of China
Chengdu, China
ruijinwang@uestc.edu.cn

***Abstract*—The rise of big data age in the Internet has led to the explosive growth of data size. However, trust issue has become the biggest problem of big data, leading to the difficulty in data safe circulation and industry development. The blockchain technology provides a new solution to this problem by combining non-tampering, traceable features with smart contracts that automatically execute default instructions. In this paper, we present a credible big data sharing model based on blockchain technology and smart contract to ensure the safe circulation of data resources.**

Keywords-big data; blockchain; smart contract; data sharing

I. INTRODUCTION

Since 2012 [10], the term, big data, has been mentioned more and more frequently. In the era of information explosion, massive data has greatly promoted the development of big data technology and innovation. It is reported that the total global data will be about doubled every two years, and by 2020, the total global data available is predicted to reach 35ZB [8]. The data size in this two years will be equal to the amount of data that humans generated before. Recently, the Tencent Institute released the 2016 "Internet +" index [15], based on more than 73,500,000GB user's data from Tencent's WeChat, mobile QQ and WeChat Official Accounts, which is equivalent to 800 American Library of Congress. Nowadays, big data is considered to be as precious as gold, for example, government use big data to analyze and as a indicator for making decision, and companies use big data to predict the users' interests and needs. According to the McKinsey Global Institute report [1], retailers have increased the potential operating margins by 60% possible with big data. It clearly shows that big data has become part of economic growth.

However, in the era of big data explosion, the risk of data circulation and the trust issue of data source has gradually attracted public attention. In October, 2016, the "DDos attack", which swept through the United States [2], has caused great panic in data security. The defect of centralized storage of data resources was revealed [14]. At the point of data circulation, the two sides of the data transaction do not trust each other, that the parties are always suspicious about the data sharing: government in the open data is cautious about data sharing, fear of violation in national security; enterprises take data resources as important assets, and are not willing to open the resources; individuals have privacy concerns about personal information. The intervention of multi-parties has led to poor control over data transfer, which obviously influence the circulation of data and hinder the further development of the entire big data industry.

Related Work. Currently, the solution to this issue is more based on legal and moral aspects. In 2016, European Union passed the General Data Protection Regulations [3]. In April the same year, China Academy of Information and Communications Technology (CAICT) issued the Data Security Circulation Convention. These regulations promote the standardization of big data industry, which protect the data circulation to a certain degree. However, as the behavior of each person cannot be completely controlled only from legal and moral aspects, this problem is not fundamentally solved.

But with the arrival of the blockchain 2.0 era [9], solving the problem of big data security circulation has become possible. Blockchain 2.0 introduces the concept of smart contracts [12], which is no longer limited to transactions between currencies, and there will be more extensive instruction embedded in the blockchain. The smart contract does not need mutual trust, as it is not only defined by the code, but executed by the code. Besides, it's completely automatic and cannot be intervened.

1) Blockchain storage model, with non-tampering feature and traceability, ensures the privacy and credibility of the data. 2) The smart contract that automatically execute the default instruction and the complete de-centric model guarantee the security of data sharing. 3) Establish a reliable big data distribution system without trusting third parties.

Organization. Section □ introduces the existing problems of big data and our solution. Section □ introduces the design and implementation of the model. Section □ is the security analysis of the model. Section □ is the summary and prospect of this paper.

II. SITUATIONAL ANALYSIS

Data source is the core of the development of big data, and currently, there exists a lot of unavoidable problems in big data circulation. Data concentrated in government agencies, Internet giants, communications companies, and they master the first-hand data, which are of high value, good quality,

978-1-5386-3349-6/17 $31.00 © 2017 IEEE
DOI 10.1109/BIGCOM.2017.31

rapid growth, etc. In the exploration of data resources, big data research companies want to get more and better data sources for deep mining, while the parties are hesitant to share the data [13]. For example, governments are worried about negative outcomes relating to confidence after sharing data; and for enterprises, the data represents the assets which cannot be easily shared. As a consequence, data “island” is formed, and for those data-driven companies, they are stuck with the situation because there is no good quality data support and data credibility is not conducive to scientific research, which finally affecting the development of big data industry.

In this paper, we will introduce the blockchain technology to solve the above problems, blockchain and smart contract are combined to build a reliable data sharing model without the reliable third party, breaking the current data "island" and improve data credibility.

Data Storage. The parties that have the data source use the blockchain as a data storage scheme, data is placed in a specific way to link to the block. Timestamp server, encryption algorithm and Proof-of-Work ensure data confidentiality and credibility.

Safe Sharing. Aggregate the owner of the major data sources together to build a blockchain network, and each of them serve as the network node to maintain the contents of the block. Blockchain information synchronizes between the various nodes and ensure the safety of data sharing.

Authority Management. The data owner authorizes the data to the data requester. Automated execution of the smart contract ensures the entire process to be open and transparent. Authorized owner can get the data sources from blockchain which has a quality assurance of data sources.

III. Design and Implementation of Model

This section mainly introduces the architecture design and implementation of big data blockchain network.

A. Architecture Design

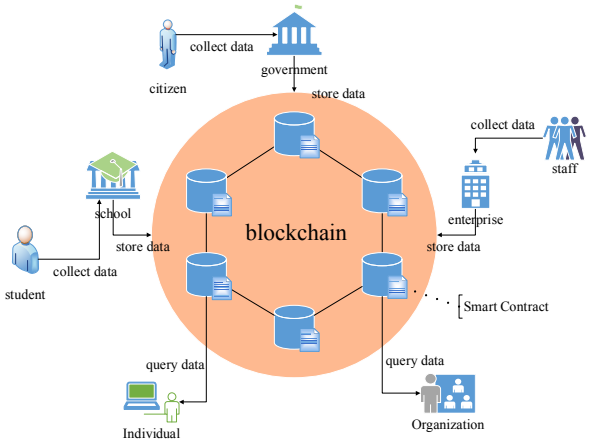

Figure 1. Overview of the architecture of platform.

As shown in Fig. 1, our system design three major parts, namely, data producers, blockchain network and data users.

1) The data producers are divided into school, government and business. With the school collecting students' data, the government collecting the citizens' data, and the enterprise collecting the users' data. These three parties will collect the massive data through the smart contract to store in the blockchain, for the use of data sharing.

2) The smart contract code runs on the contract layer of blockchain, which provides the authority to control the system. Each data producer corresponds to deploying a smart contract on the blockchain and storing the data through it.

3) The data consumers are also divided into individuals who need data assets or organizations that conduct research on big data. They can initiate a data transaction request, which requests the data producer to provide data usage rights, and obtain an authorized data set from the blockchain through a smart contract.

As it shows in Fig. 1, we designed the smart contract which only accept two types of request: one is to store data requests, we call it T_s; the other is the query data request, we call it T_q. Please imagine a scenario: Enterprise A buys big data sets from enterprise B through our platform. Assuming that each enterprise has only one blockchain node (in fact, there can be multiple), since each node of the blockchain network has a private key control address, enterprise A only needs to know the address of Node B. Through the deployed smart contract S, enterprise A initiates a request for a T_q type to Node B, synchronizes the blocks through the confirmation of the miners' nodes and links them to the blockchain. Node B authorizes A, and then enterprise A can obtain the purchased data set from the blockchain. Because the data stored on the blockchain has a non-tampering feature, enterprise A does not have to worry about the risk of being tampered with the data set purchased. And the data on the block chain is stored by asymmetric encryption, and only authorized objects can access the data, so Enterprise B does not have to worry about the data set will be leaked.

B. Data Storage Protocol Based on Blockchain

The data source is stored in the blockchain, and the blockchain is the distributed data account book that each node shares. The transaction information and authentication information in the block are transparent. The data information is stored after encrypted by private key and can be accessed and authorized by the data owner.

The Construction of Data Blocks. The miners encapsulate the transaction data and code received over a period of time into a data block with timestamps through a specific hash algorithm and MPT operation. The irreversibility of time and the unidirectionality of the hash value guarantee the authority of the data.

The Data Block Contains Two Parts. The block header and the block body. The block information is divided into three categories: H is the collection of the relevant information of the block, T represents the corresponding transaction information of the block, and U represents the other block information contained in the block [4]. Therefore, a block B is defined using the equal relationship as Eq. (1).

$$B \equiv (B_H, B_T, B_U) \tag{1}$$

The detailed block data structure is defined as follows.

Prototype 1 Data Block Structure

```
header     *Header  // block header
uncles     []*Header // uncle block information
transactions  Transactions // Recording trade information
```

Proof-of-Work. In the big data blockchain network, the miners' node collects near transactions, by continually trying random numbers [5]. This random number is valid if the output of the final algorithm reaches the target we had previously set. In this way the miners' workload is guaranteed, the DDOS attack is resisted and will output 256-bit digits. Formal description is as Eq. (2).

$$m = H_m \cap n \leq \frac{2^{256}}{H_d} \; with \;\; (m,n) = PoW(H_{\neq n},\, H_n, d) \quad (2)$$

Where $H_{\neq n}$ is the new block's header hash without the nonce and mix-hash components. H_{d} is the di□culty value. PoW is the Proof-of-work function as follows.

Algorithm 1 Proof-of-Work

Input: $H_{\neq n}$ is new block's header hash without nonce, d is difficulty value

```
procedure proof_of_work(H_{≠n}, H_n, d):
    procedure sh(h,n):
        return KEC512(h+n[||n|| − i])
    end procedure
    procedure mixhash(H_{≠n}, H_n, d)
        procedure mixdataset(d,m,s,i):
            temp = FNV(i ⊕ s[0], m ⌈i mod ⌊mixbyte/wordbyte⌋⌉)
            newdata(d,m,s,i)[j]=d[ temp mod ⌊size/hashbyte/mix⌋ ∗ mix + j]
            return FNV(m, newdata(d,m,s,i))
        end procedure
        procedure compress(m,i):
            if(i ≥ ||m||-8) then:
                return m
            else  compress i+4 to i+8
        end procedure
        h = KEC(RLP(L_H(H_{≠n})))
        if(i==J_accesses − 2) then:
            access(d,m,s,i) = mixdataset(d,m,s,i)
        else access(d,m,s,i) = access(mixdataset(d,m,s,i),s,i+1)
        return compress(access(d, ∑_{i=0}^{mix} sh(h,n), −1), −4)
    end procedure
    m = mixhash(H_{≠n}, H_n, d)
    n = KEC(Sh(H_{≠n}, H_n)+mixhash(H_{≠n}, H_n, d))
    return (m,n)
end procedure
```

Consensus Algorithm. The miners' node successfully generates block in the above way, and the block are waiting for confirmation of other miners' nodes in the blockchain network. The system specifies that the block need to be verified by more than half of the nodes. Then the block will be linked to the existing blockchain. We define the verification function as follows.

Algorithm 2 Block Validation

Input: *block* is the object of new block

```
procedure proof_valid_of_block(block):
    if V(H) ≡ Math.pow(2,256) / H_d && m = H_m && H_d = D(H):
    if H_g ≤ H_1 &&
        H_1 > P(H)_{H_1} + ⌈P(H)_{H_1}/1024⌉ && H_1 > P(H)_{H_1} - ⌈P(H)_{H_1}/1024⌉ then:
        if H_1 ≥ 125000 && H_s > P(H)_{H_s} && H_1 = P(H)_{H_1} + 1:
            if ||H_x|| ≤ 32 then:
                return true;
    return false;
end procedure
```

C. Data Circulation Protocol Based on Smart Contract

The smart contract is the authoritative guarantee of the security of big data circulation. It is a computerized transaction protocol, which is completely automatic and no longer needs to be supervised [7]. In the contract layer of the blockchain network, the smart contract is a set of scenario-based Procedural rules and logic. In the Blockchain, network transactions are divided into two, one is to create a contract, and the other is generating message call through the smart contract. Smart contract operation model is shown in Fig. 2.

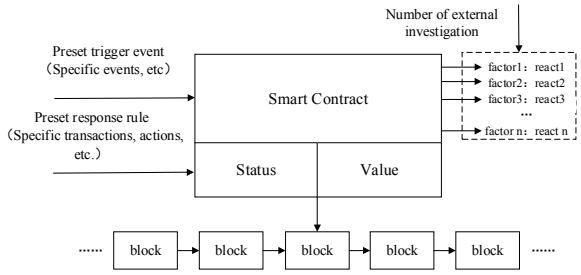

Figure 2. smart contract operation model.

The smart contract is attached to the blockchain in the form of program code, which is recorded in the specific block of the blockchain after distributed network propagation and node verification. The smart contract encapsulates several predefined state and conversion rules, triggering the execution of the contract (such as reaching a specific time or a specific event, etc.), responding to a particular scenario, and so on. At the same time, the blockchain network can monitor the status of the smart contract in real time and execute the contract by verifying the external data source and confirming that the specific trigger condition is satisfied.

Construction of Data Circulation Protocol. The data sharing blockchain network is designed with smart contract-driven data circulation protocol, which is used to invoke the contract interface for data storage and authorization management. The circulation protocol is constructed as follows.

When the data producer joins the big data blockchain network, it automatically initiates a transaction for creating contract. The address of the contract corresponds to the

address of the data producer. The data producer can store the data in the blockchain by calling the store interface of the contract. The data demander initiates a data authorization request to the contract owner through the contract address and the abi call query interface. The data owner authorizes the data through *addAuthorize* function. The data circulation protocol built through the smart contract ensures the safe sharing of data.

Protocol 1 Data Circulation

```
address private owner;
mapping (bytes32 => string) private data;
mapping (address => bool) private AuthoList;
function big_data(){ owner = msg.sender; } // constructor
function store(bytes32 key,string value){ // store data
    if(msg.sender != owner) throw;
    data[key] = value;
}
function query(bytes32 key) constant returns(string){ // query data
    bool autho = AuthoList[msg.sender];
    if(autho == true){
        return data[key];
    }
}
function addAuthorize(address user){ // manage authority list
    if(msg.sender != owner) throw;
    AuthoList[user] = true;
}
```

Data Authorization Management. As the Fig. 3 shows, when the data producer A needs to share data to the data demander B, the first thing is to reach a consensus with B to develop constraints (data range, aging, etc.). Then the smart contract use the public key of A Puk_A to decrypt the data and output the corresponding result according to the constraint. At the same time , A needs add Puk_B to the authorization list, After using Puk_B to encrypt the data, the smart contract outputs data to the B. And after B use the private key Prk_B decrypt dataset, A and B successfully share data.

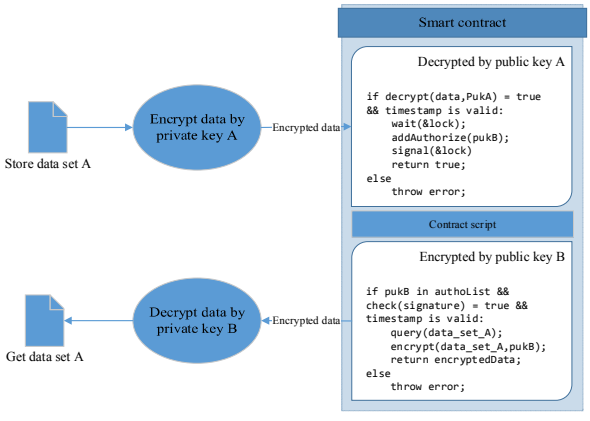

Figure 3. data authorization management model.

IV. SECURITY ANALYSIS

This section describes in terms of architectural security and data security in blockchain.

A. *Architectural security*

The current Internet platform is built on the structure of the central server. Traditional relational database leads to the over-centralization of data storage, causing the security of the data greatly depends on the central server security. The majority of data leaks result from hacker attacks on the server, such as DDOS attacks, SQL injection attacks, CC attacks [11].

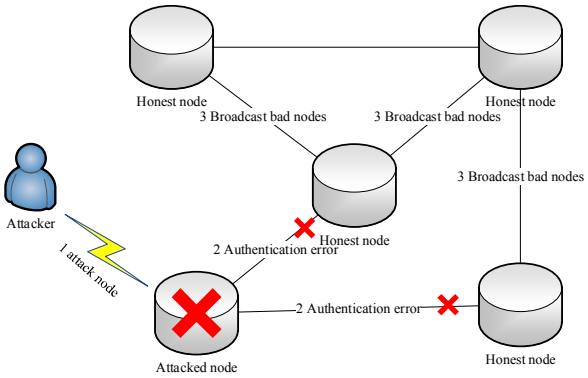

Figure 4. Simulation Attack.

As shown in Fig. 4, the big data de-centered storage mode allows the block information to be maintained by nodes in the network. Hackers can control only few node in the network, but the data in the blockchain encrypted through the private key, which ensures the confidentiality of the data. If hackers tamper with the block data, the network's consensus model will ensure that other nodes reject the bad node. This model has resisted all attacks against traditional data storage patterns.

B. *Forged Block Attack Control*

Blockchain is the guarantee of data security. Assuming that there is an attacker serving as a bad node in blockchain network, forging blocks to link. The competition between the honest chain and the chain of attackers can be described by random walking with a binary tree [6]. The probability of success of the forged block attack is calculated as Eq. (3).

$$P_Z = 1 - \sum_{k=0}^{z} \frac{\lambda^k e^{-\lambda}}{k!} * (1 - (\frac{q}{p})^{z-k}), \quad \lambda = Z\frac{q}{p} \qquad (3)$$

p is the probability of honest nodes to generate the next block, q is the probability of the attacker to generate the next block, P_z is the probability of the attacker will ever catch up the main chain from z blocks behind. The algorithm is designed as follows.

Algorithm 3 Forged Block Attack Success Rate

```
Input: q is probability of attacker to generate next block
       z is catch up the main chain from z blocks
procedure attackerSuccessProbability(double q, int z):
    double p = 1.0 – q;
```

```
3:      double lambda = z * (q / p);
4:      double sum = 1.0;
5:      for(k = 0;k <= z; k++)
6:          double poison = exp(-lambda);
7:          for(i = 1;i <= k; i++)
8:              poison *= lambda / i;
9:          sum -= poison * (1- pow(q / p, z - k));
10:     return sum;
11: end procedure
```

By calculating the value of P_z for q = 0.1 and q = 0.3, the statistical results are shown in Fig. 5.

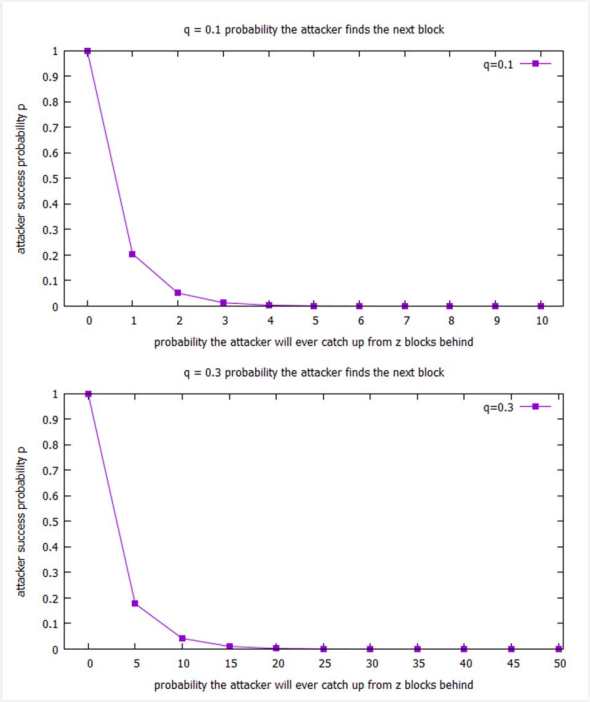

Figure 5. attacker success probability.

Through the chart above, we can find that the value of z exponentially declines. There are a large number of nodes and computing power in blockchain network, so the attacker needs more computing power than the entire network has to create a block and modify all nodes' record in blockchain, which is almost impossible.

V. CONCLUSION

In the era of big data, the aggregation of data sources allow big data to play a valuable role. However, the existing big data trust system is not perfect, and the security of data circulation and the availability of data sources cannot be guaranteed, resulting in data "island" problem. We provide a credible data-sharing platform for data producers and demand parties by building a decentralized data circulation security system based on blockchain and smart contract. Blockchain ensures data traceability, and the automated execution of the smart contract provides protection for data security sharing.

For the data provider, the decentralized architecture avoids the data security risks brought by the over centralization of data storage; As for users, the blockchain operation model ensures the transparency in the process of information collection, and brings stronger protection for users' privacy and the right to know.

ACKNOWLEDGMENT

This paper was supported in part by following funds: Science and Technology projects in Sichuan Province (2016ZC2575, 2015JY0178, 2014GZ0109, 2015KZ002, 2015JY0030).

REFERENCES

[1] Manyika J, Chui M, Brown B, et al. Big Data: The Next Frontier For Innovation, Competition, And Productivity[J]. Analytics, 2011J.

[2] Jin D, Hannon C, Li Z, et al. Smart street lighting system: A platform for innovative smart city applications and a new frontier for cyber-security[J]. The Electricity Journal, 2016, 29(10): 28-35.

[3] Tsakalakis, Niko, S. Stallabourdillon, and K. O'Hara. "What's in a name: the conflicting views of pseudonymisation under eIDAS and the General Data Protection Regulation." European Journal of Psychotraumatology3.2(2016):163-167.

[4] Buterin V. A next-generation smart contract and decentralized application platform[J]. white paper, 2014.

[5] Wood G. Ethereum: A secure decentralised generalised transaction ledger[J]. Ethereum Project Yellow Paper, 2014, 151.

[6] Nakamoto S. Bitcoin: A peer-to-peer electronic cash system[J]. 2008.

[7] Norta A. Creation of smart-contracting collaborations for decentralized autonomous organizations[C]//International Conference on Business Informatics Research. Springer International Publishing, 2015: 3-17.

[8] Aljohani N R, Davis H C. Learning analytics in mobile and ubiquitous learning environments[J]. 2012.

[9] Swan M. Blockchain: Blueprint for a new economy[M]. " O'Reilly Media, Inc.", 2015.

[10] Lohr S. The age of big data[J]. New York Times, 2012, 11(2012).

[11] Simmons C, Ellis C, Shiva S, et al. AVOIDIT: A cyber attack taxonomy[J]. 2009.

[12] Peters G W, Panayi E. Understanding Modern Banking Ledgers through Blockchain Technologies: Future of Transaction Processing and Smart Contracts on the Internet of Money[M]//Banking Beyond Banks and Money. Springer International Publishing, 2016: 239-278.

[13] Kitchin R. The data revolution: Big data, open data, data infrastructures and their consequences[M]. Sage, 2014.

[14] Boyd D, Crawford K. Six provocations for big data[C]//A decade in internet time: Symposium on the dynamics of the internet and society. Oxford: Oxford Internet Institute, 2011, 21.

[15] 佚名. 中国“互联网+”指数(2016)发布腾讯研究院深度解读数字中国[J]. 新经济导刊, 2016(8):78-78.