Research on Big Data Storage Method based on IPFS and Blockchain

Jing Tang
China Unicom System Integration Limited Corporation, China

Tao Jia*
China Unicom System Integration Limited Corporation, China

Haibo Chen
China Unicom System Integration Limited Corporation, China

Chuncheng Wei
China Unicom System Integration Limited Corporation, China

ABSTRACT

With the popularity of digital cryptocurrency such as bitcoin, blockchain, as a new distributed framework with decentralization, non rewriting and traceability, has sprung up rapidly and has been applied in many industries such as finance, medical treatment, information security, etc. In order to ensure the security of transaction data, all key information in the business needs to enter the blockchain network. In the field of artificial intelligence, model data (i.e. effective feature point data set) will be the key information and will be used frequently. However, these feature point datasets may be megabytes, auspicious, or even terahertz. Then, the performance of big data transaction in blockchain network will be a problem worthy of study. Therefore, this paper proposes a blockchain big data storage method based on IPFs. This method mainly solves the transaction performance problem of large text data in the blockchain network. The data larger than 100 megabytes are stored in IPFs to obtain the hash certificate of text. The hash code is the only transaction voucher in the blockchain network. It greatly improves the transaction efficiency of blockchain network. In this paper, a comparative experiment is set up to further prove the efficiency of our method.

CCS CONCEPTS

• **Computing methodologies**; • **Modeling and simulation**; • **Machine learning**;

KEYWORDS

IPFS, blockchain, BigData, Decentralization

ACM Reference Format:
Jing Tang, Tao Jia, Haibo Chen, and Chuncheng Wei. 2020. Research on Big Data Storage Method based on IPFS and Blockchain. In *2020 2nd International Conference on Video, Signal and Image Processing (VSIP '20), December 04–06, 2020, Jakarta, Indonesia.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3442705.3442714

*Tao Jia is the corresponding author.

VSIP '20, December 04–06, 2020, Jakarta, Indonesia

ACM ISBN 978-1-4503-8893-1/20/12…$15.00
https://doi.org/10.1145/3442705.3442714

1 INTRODUCTION

Blockchain uses chain data structure to verify stored data, and uses distributed node consensus algorithm to generate and update data. Using cryptography to ensure data security. It can be said to be a safe, stable and trusted technology [1–5]. With the more mature of the technology, at present, blockchain has shown great influence in the field of information technology such as big data, Internet of things, artificial intelligence, etc., and the information transmission performance of blockchain network is required to be higher and higher. It can be divided into private alliance and blockchain. The public chain is completely decentralized. The nodes in the distributed system can participate in the operation, verification and consensus of the data on the chain. The alliance chain is partially decentralized, which is suitable for the alliance composed of multiple entities. The private chain is completely decentralized. Among them, alliance chain has more advantages in efficiency and security. This paper also focuses on the optimization of alliance chain. At present, the general blockchain network does not have high-speed processing of large-scale text data, and its performance is low, and it can process less than 10 transactions per second. In order to solve this problem, this paper proposes a blockchain big data storage method based on IPFS. The interplanetary file system (IPFS) is introduced to store large-scale text to generate file hash code [15–17], which is used in blockchain network transactions. For a large number of large-scale blockchain related services, the method proposed in this paper can effectively provide the operation efficiency of the network and greatly reduce the redundant transmission of the network. Finally, the feasibility of our method is further proved by experiments.

2 METHODOLOGY

2.1 The Technology of Blockchain

Blockchain is composed of a series of block links that record transaction data generated according to time sequence, forming an open and trusted transaction database [6–10]. Any transaction has a complete evidence chain and trust traceability link. Blockchain technology is based on the elliptic curve digital signature algorithm in cryptography to realize the centralized point-to-point design. Its technical principle is shown in Figure 1. The blockchain consists of a series of blocks arranged in chronological order. Each block is composed of the hash value of the previous block and the content, time stamp, data signature and consensus mechanism of the block [11–14, 21–24].

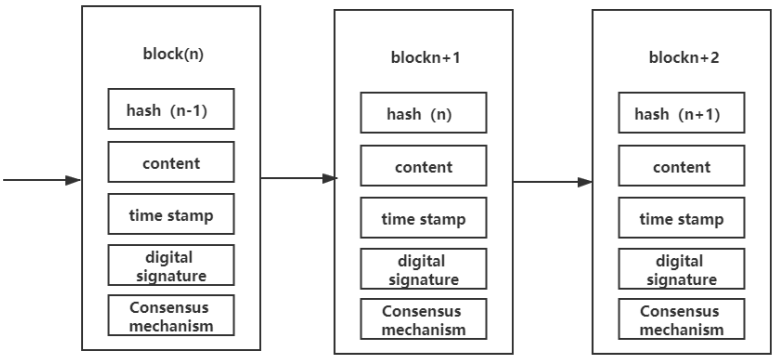

Figure 1: The Principle of blockchain Technology

The blockchain network designed in this paper is composed of CA, the management of organization, the management of node, the management of alliance, the management of channel and smart contract.CA: With fabric CA mechanism and national certification as mechanisms, account authentication and node authentication of organizations in the alliance are realized; the management of organization: Users can set up organizations. In this paper, we refer to the projects that we need to link up. the management of node: n order to ensure the effectiveness of the transaction, it needs to broadcast to the network through nodes, so it is necessary to manage many nodes in a unified way. the management of alliance: Multiple channel chains can be established in the alliance, and multiple smart contracts can be run in the channel, which has a perfect member invitation mechanism. the management of channel: The alliance establishes channels according to the business logic between organizations, and channels will not interfere with each other. smart contract: An agreement to disseminate, validate, or execute contracts by means of information.

Blockchain applications interact either directly with blockchains or with smart contracts in order to achieve consensus on transactions, data or code execution. But there are many problems when working with large data files. Because the files are usually not required for the blockchain nodes to function, the blockchain becomes bloated, resulting in data being replicated on a large amount of nodes. First, storing large files on the blockchain is inefficient. Limitations on the block size require files to be split and reassembled off-blockchain. Additional data relevant to reassembling files would also have to be stored, requiring either even more space or a distinct system that provides the reassembly information. If the smart contract is used to directly store the file part, the data can be accessed more directly and conveniently, and the important data can also be stored. However, the smart contract is very expensive and needs to do data synchronization and node verification on each node. Secondly, operating the nodes becomes more expensive. More data needs to be propagated through the network, processed and stored by the node. Nodes would thus require connections with higher bandwidths and more storage space to store the blockchain, The end result is a straight-line increase in costs [18–20].

It concludes that blockchains are not the right platform to share and store large files. Fortunately, the Interstellar file system can be leveraged to support applications while keeping the blockchain small in size. A decentralized access model, with privacy built into it by design, is made possible by a software stack of blockchains working on top of a decentralized peer-to-peer file system. Users can efficiently share large files and still benefit from the blockchain. Cryptographic hashes that serve to securely identify a file's content, can be send to the latter, thus proving that the file was available to someone at a certain time [28–33].

2.2 The Interplanetary File System

The Interstellar file system is a decentralized storage network based on blockchain. The following example illustrates how IPFS functions. When a file is uploaded to IPFS, it is split into chunks5, each containing at most 256 kilobytes of data and/or links to other chunks. Every chunk is identified by a cryptographic hash, also named content identifier, that is computed from its content. The mentioned links also contain content identifiers, thus forming a Merkle directed acyclic graph (Merkle DAG) that describes the file as a whole and can be used to reconstruct any file from its chunks. Because of the Merkle DAG, an entire file can be identified by just using the root hash. Once a node has divided the file into chunks, and the Merkle DAG has been formed, the node registers itself as a provider by means of the DHT. The DHT is essentially a distributed key-value store. it uses node identifiers and keys – both need to have the same length - together with a distance metric to easily store and retrieve information. When looking for a value, a node attempts to find nodes that are close to the key and requests the value from them. It does that by using buckets to keep track of nodes within the network. The buckets are organized such that any node of the network has precise information of its near environment. However, a node's knowledge of other nodes decreases as the distance increases. Thus, in order to find the value associated with a key, a node contacts a node that is closer to the key than it is itself. The latter replies either by returning the value, or by referring to nodes that are even closer to the key. This continues until the key is found [28–33]. Storage performance: it is different from the traditional network, using HTTP protocol, but through content addressing technology to locate resources. It stores multiple files on different computers through hash calculation. Users can

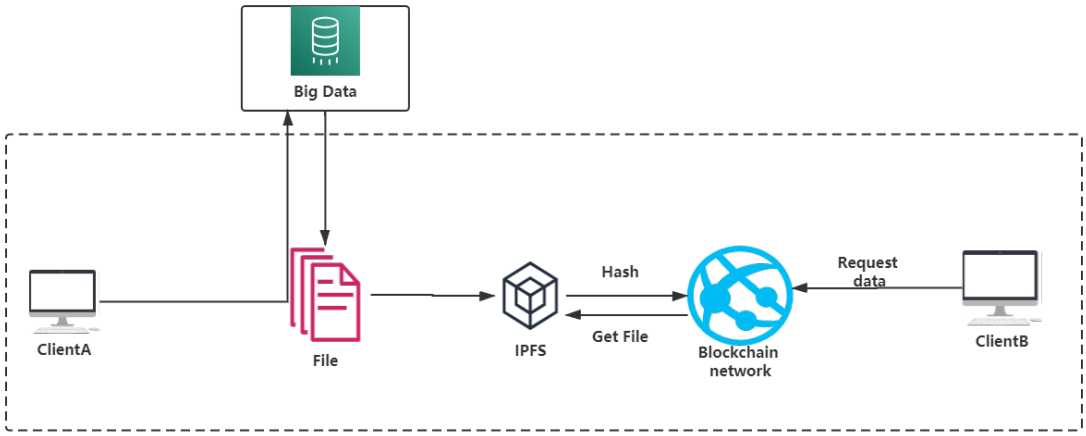

Figure 2: The Big data storage method based on IPFS

access a file according to the hash address. In addition, it has the peer-to-peer feature. It stores the files in the computers or servers closest to the users, and the loading speed is greatly improved. Data security: the stored files are distributed in multiple servers, and the files are more secure to avoid file loss caused by server crash, interruption and theft. Data redundancy: using protocol to solve the problem of content redundancy, optimize the storage of duplicate files, and save resources.

2.3 Big Data Storage Method based on IPFS and Blockchain

The big data storage method based on IPFs proposed in this paper mainly aims at optimizing the transmission performance of large text information in blockchain network. In order to provide network efficiency, the large text data is divided into blocks: by date, by function point, by data size, and so on. The specific granularity is defined according to their own business requirements. Block data is stored in IPFS to obtain the hash code corresponding to the text. The IPFS can deal with data redundancy and optimize the characteristics of duplicate files, which greatly saves resources. The hash code obtained from IPFS will be sent to the business transaction of blockchain network as the unique document of text. After the end of each transaction, the hash code is used to obtain the corresponding data block from IPFS. We use this strategy to improve the transaction efficiency of blockchain network and save storage resources. The Big data storage method based on IPFS is shown in Figure 2

3 EXPERIMENTS

In order to verify the effectiveness of this method, we set up a set of comparative experiments. The data set adopts cord-19, which is popular this year, with a size of about 4.25g, the White House and a coalition of leading research groups have prepared the COVID-19 Open Research Dataset (CORD-19). This freely available dataset is provided to the global research community to apply recent advances in natural language processing and other AI techniques to generate new insights in support of the ongoing fight against this infectious disease. The experimental design adopts the idea of generating confrontation network game, sets up the cord model network and the cord verification network, establishes the channel, and joins the two networks into the channel, and formulates the contract of the two networks trading in the blockchain. The network component diagram of this experiment is shown in Figure 3

Table 1: The number of transactions completed in the fixed time of the two groups of experiments

Method	1(s)	5(s)	10(s)
Netword	9	46	95
IPFS-Netword	50	300	600

In the data preprocessing stage, we select the target data from January to October in 19 years as the research object, and divide the target data into 30 data blocks according to the time, and generate one data block every 8 days on average. Next, build our blockchain network (refer to Section 2.1), and the cord model network and the cord verification network are also on the chain. When the network generates data, the data is encrypted and sent to the verification node. If the verification node identifies any unauthorized network incoming packets, the verification program will block the A node introduces the data of this network. If the verification node passes the check authentication, then it is stored in the IPFs or updated IPFs file, and its hash value is stored in the smart contract.

Set two groups of experiments of IPFs based blockchain network (IPFs netword) and non IPFs storage blockchain network (netword) as a comparison. Table 1 shows the number of transactions completed in the fixed time of the two groups of experiments.

We can see that the method proposed in this paper can significantly improve the transaction efficiency of the blockchain network, and further verify the feasibility and superiority of the blockchain big data storage method based on IPFs proposed in this paper.

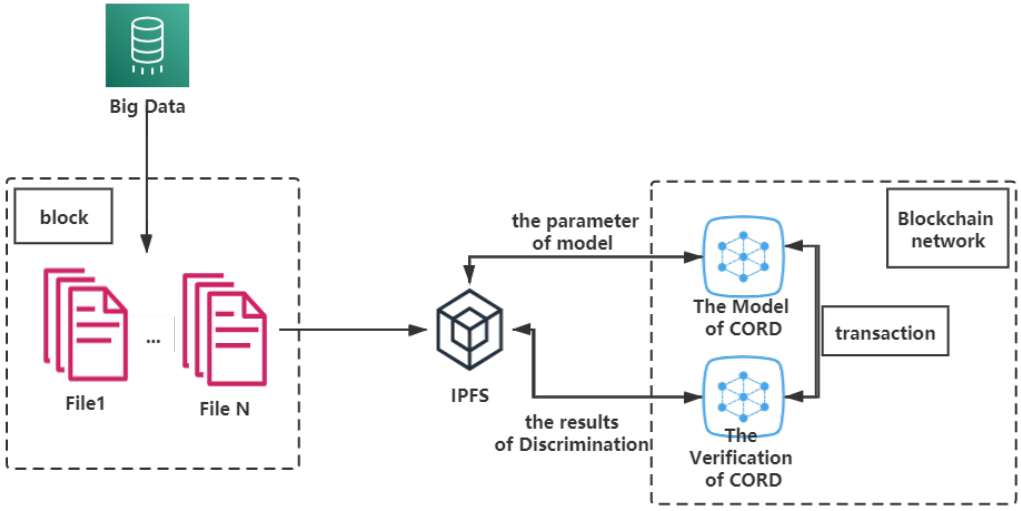

Figure 3: Component diagram of cord experimental network

4 APPLICATION

4.1 Medical Compensation

In clinical trials, the use of IPFs based blockchain can help to eliminate the forgery of data and the omission of adverse results of clinical research. Because of the anonymity of the inherent coding of data, blockchain makes it easier for patients to grant Data permission for clinical trials, In addition, the tamper proof features of blockchain can prove the integrity of the data collected through the blockchain for clinical research. The transparency and commonality of blockchain also make the research on data replication based on blockchain easier, which will make it possible for blockchain to completely change biomedical research.

Therefore, we use hyperledger fabric to build alliance chain, and its data storage is distributed. In this way, all parties in the distributed ledger can be synchronized with each other in the blockchain. Different from the self managed database, blockchain provides an independent data sharing platform, which can be used by different hospitals for cross chain merging and intelligent aggregation according to different chains or data formats used by different hospitals. Data aggregation between patients and hospitals is shown in Figure 4

Our alliance chain has a special endorsement policy mechanism. The endorsement policy can be written according to the business logic. After verification by each node, endorsement documents can be formed. The scope of application includes CT, PDF and other supporting documents;

After the file is stored by IPFS, it is converted into hash code. Hash code is used to verify in each node of the platform, which can be used as the certificate of medical insurance or insurance. Paperless information sharing, reduce notarization business errands, letter plus letter, transparent and rigorous [24–27].

Table 2: The number of transactions completed in the fixed time of the medical compensation

Method	1(s)	5(s)	10(s)
Netword	20	67	200
IPFS-Netword	110	450	1003

Based on the medical insurance certificate and insurance certificate in the chain, the insurance company continues the reimbursement process for patients. Reimbursement process based on IPFS alliance chain is shown in Figure 5

Because of its decentralized, unchangeable and transparent features, our alliance chain blockchain provides a trusted platform for patients, hospitals and insurance companies.

We only apply the method proposed in this paper to the medical reimbursement scenario. In addition, we can also apply it to the identity authentication, which can authenticate the medical records, medical records, doctors and insurance company personnel with digital identity, so as to protect the agency right of each patient and the right of others to access their medical record data. The successful implementation of the case scenario further demonstrates the potential of the proposed method to change the peer review process of clinical research publications.

The function advantage of our method network has been obvious. How about the network performance parameters? Here we continue to use the method of the experimental part of this paper, using the number of transactions per unit time of the network to measure the performance of the network. The number of transactions completed in the fixed time of the medical compensation is shown in Table 2

5 FUTURE WORK

We aim to continue to research on big data processing through blockchain technology, we plan to develop a larger scale blockchain

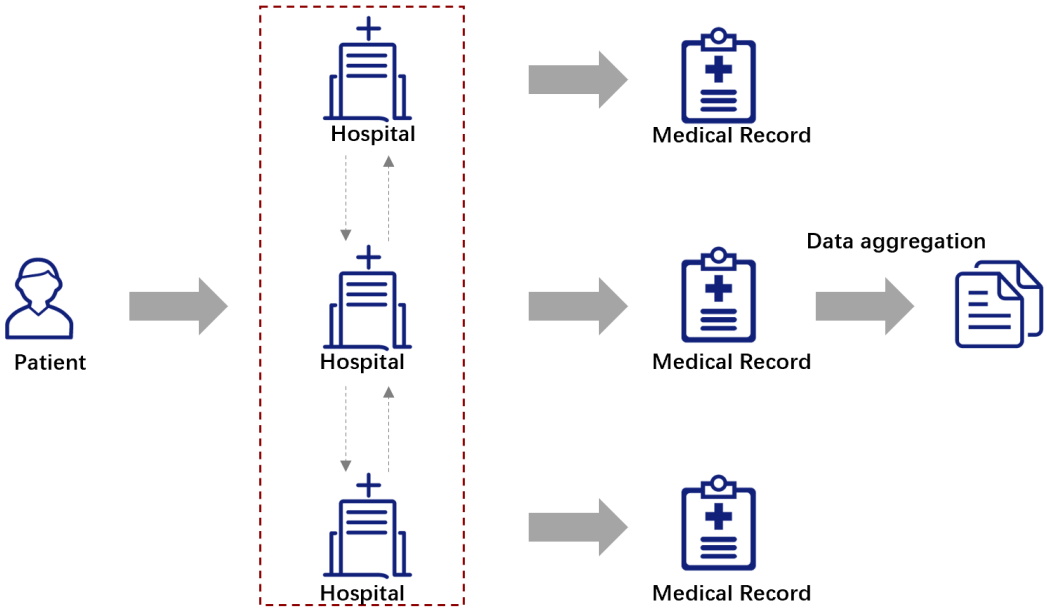

Figure 4: Data aggregation between patients and hospitals

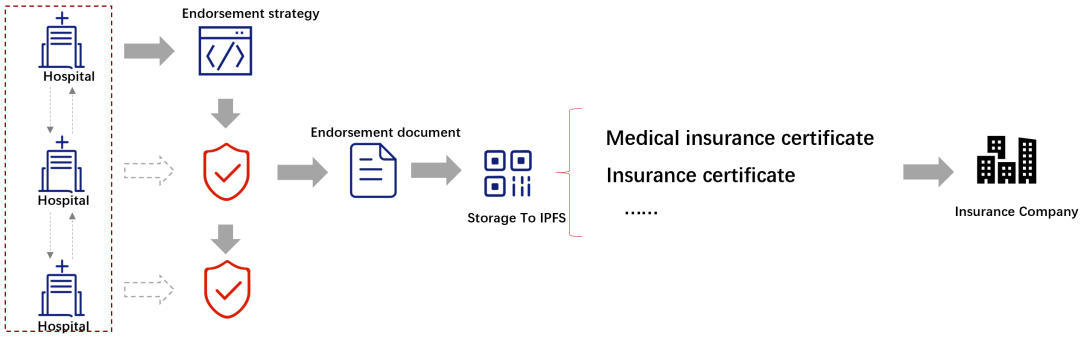

Figure 5: Reimbursement process based on IPFs alliance chain

development platform and introduce hyperledger fabric 2.0 into our structure to optimize the current structure performance. In addition, kubernetes is further planned to optimize the deployment process, improve the efficiency of network deployment, and combine with containerization technology to improve the efficiency of blockchain network. More work will be done on the data flow transmission efficiency and user identity authentication within the blockchain network.

6 CONCLUSIONS

In order to solve the problem that the general blockchain network does not have high-speed processing of large-scale text data, and the performance is low. This paper proposes a blockchain big data storage method based on IPFs. In order to verify the feasibility of the method, we design the experimental scenario of the trade between the cord model and the cord verification model. Compared with the non IPFS blockchain network, our method significantly improves the transaction efficiency of the blockchain network, and has obvious advantages.

At present, we have applied this method to the medical reimbursement scenario. We store the patient's reimbursement voucher, insurance certificate and other related file information into IPFS, and verify and trade the obtained Hash code at each node of our alliance chain, so as to reduce the notarization business process and be transparent and rigorous.

Based on the current research results, we hope that the research results will be more applied to the field of neural network.

REFERENCES

[1] Li X, Jiang P, Chen T, *et al.* A survey on the security of blockchain systems[J]. Future Generation Computer Systems, 2020, 107: 841-853.

[2] Taylor P J, Dargahi T, Dehghantanha A, *et al.* A systematic literature review of blockchain cyber security[J]. Digital Communications and Networks, 2020, 6(2): 147-156.

[3] Joshi G P, Perumal E, Shankar K, *et al.* Toward Blockchain-Enabled Privacy-Preserving Data Transmission in Cluster-Based Vehicular Networks[J]. Electronics, 2020, 9(9): 1358.

[4] Xiao Y, Zhang N, Lou W, *et al.* A survey of distributed consensus protocols for blockchain networks[J]. IEEE Communications Surveys & Tutorials, 2020, 22(2): 1432-1465.

[5] Yu H, Nikolić I, Hou R, *et al.* Ohie: Blockchain scaling made simple[C]//2020 IEEE Symposium on Security and Privacy (SP). IEEE, 2020: 90-105.

[6] Singh S K, Rathore S, Park J H. Blockiotintelligence: A blockchain-enabled intelligent IoT architecture with artificial intelligence[J]. Future Generation Computer Systems, 2020, 110: 721-743.

[7] Sharma P, Jindal R, Borah M D. Blockchain technology for cloud storage: A systematic literature review[J]. ACM Computing Surveys (CSUR), 2020, 53(4): 1-32.

[8] Xu Y, Huang Y. Segment blockchain: A size reduced storage mechanism for blockchain[J]. IEEE Access, 2020, 8: 17434-17441.

[9] Liang W, Fan Y, Li K C, *et al.* Secure data storage and recovery in industrial blockchain network environments[J]. IEEE Transactions on Industrial Informatics, 2020.

[10] Liang W, Fan Y, Li K C, *et al.* Secure data storage and recovery in industrial blockchain network environments[J]. IEEE Transactions on Industrial Informatics, 2020.

[11] Rathi V K, Chaudhary V, Rajput N K, *et al.* A Blockchain-Enabled Multi Domain Edge Computing Orchestrator[J]. IEEE Internet of Things Magazine, 2020, 3(2): 30-36.

[12] Bao J, He D, Luo M, *et al.* A survey of blockchain applications in the energy sector[J]. IEEE Systems Journal, 2020.

[13] Javed M U, Javaid N, Aldegheishem A, *et al.* Scheduling Charging of Electric Vehicles in a Secured Manner by Emphasizing Cost Minimization Using Blockchain Technology and IPFS[J]. Sustainability, 2020, 12(12): 5151.

[14] Sun J, Yao X, Wang S, *et al.* Non-Repudiation Storage and Access Control Scheme of Insurance Data Based on Blockchain in IPFS[J]. IEEE Access, 2020, 8: 155145-155155.

[15] Zheng X, Lu J, Sun S, *et al.* Decentralized Industrial IoT Data Management Based on Blockchain and IPFS[C]//IFIP International Conference on Advances in Production Management Systems. Springer, Cham, 2020: 222-229.

[16] Kumar R, Marchang N, Tripathi R. Distributed Off-Chain Storage of Patient Diagnostic Reports in Healthcare System Using IPFS and Blockchain[C]//2020 International Conference on COMmunication Systems & NETworkS (COMSNETS). IEEE, 2020: 1-5.

[17] Sun J, Yao X, Wang S, *et al.* Blockchain-Based Secure Storage and Access Scheme For Electronic Medical Records in IPFS[J]. IEEE Access, 2020, 8: 59389-59401.

[18] Sun J, Yao X, Wang S, *et al.* Blockchain-Based Secure Storage and Access Scheme For Electronic Medical Records in IPFS[J]. IEEE Access, 2020, 8: 59389-59401.

[19] Krejci S, Sigwart M, Schulte S. Blockchain-and IPFS-Based Data Distribution for the Internet of Things[C]//European Conference on Service-Oriented and Cloud Computing. Springer, Cham, 2020: 177-191.

[20] ul Haque A, Ghani M S, Mahmood T. Decentralized Transfer Learning using Blockchain & IPFS for Deep Learning[C]//2020 International Conference on Information Networking (ICOIN). IEEE, 2020: 170-177.

[21] Cheema M A , Shehzad M K , Qureshi H K , *et al.* A Drone-Aided Blockchain-Based Smart Vehicular Network[J]. IEEE Transactions on Intelligent Transportation Systems, 2020.

[22] Tian Z , Zhong R Y , Barenji A V , *et al.* A blockchain-based evaluation approach for customer delivery satisfaction in sustainable urban logistics[J]. International Journal of Production Research, 2020(7):1-21.

[23] Chen B , Wu L , Wang H , *et al.* A Blockchain-Based Searchable Public-Key Encryption With Forward and Backward Privacy for Cloud-Assisted Vehicular Social Networks[J]. IEEE Transactions on Vehicular Technology, 2020, 69(6):5813-5825.

[24] Leeuwen G V , Alskaif T , Gibescu M , *et al.* An integrated blockchain-based energy management platform with bilateral trading for microgrid communities[J]. Applied Energy, 2020, 263(C):114613.

[25] Chen Y , Li H , Li K , *et al.* An improved P2P file system scheme based on IPFS and Blockchain[C]// 2017 IEEE International Conference on Big Data (Big Data). IEEE, 2017.

[26] Marjit U , Kumar P . Towards a Decentralized and Distributed Framework for Open Educational Resources based on IPFS and Blockchain[C]// 2020 International Conference on Computer Science, Engineering and Applications (ICCSEA). 2020.

[27] Kumar R , Marchang N , Tripathi R . Distributed Off-Chain Storage of Patient Diagnostic Reports in Healthcare System Using IPFS and Blockchain[C]// 2020 International Conference on COMmunication Systems & NETworkS (COMSNETS). 2020.

[28] Ali M S , Dolui K , Antonelli F . IoT data privacy via blockchains and IPFS[J]. 2017:1-7.

[29] Chen Y , Li H , Li K , *et al.* An improved P2P file system scheme based on IPFS and Blockchain[C]// 2017 IEEE International Conference on Big Data (Big Data). IEEE, 2017.

[30] Nyaletey E , Parizi R M , Zhang Q , *et al.* BlockIPFS - Blockchain-Enabled Interplanetary File System for Forensic and Trusted Data Traceability[C]// 2019 IEEE International Conference on Blockchain (Blockchain). IEEE, 2019.

[31] Zheng Q , Li Y , Chen P , *et al.* An Innovative IPFS-Based Storage Model for Blockchain[C]// 2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI). ACM, 2018.

[32] Steichen M , Fiz B , Norvill R , *et al.* Blockchain-Based, Decentralized Access Control for IPFS[C]// 2018 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData). IEEE, 2019.

[33] Steichen M , Fiz B , Norvill R , *et al.* Blockchain-Based, Decentralized Access Control for IPFS[C]// 2018 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData). IEEE, 2019.