Research Article

Shengwei Wen*

A study on the big data scientific research model and the key mechanism based on blockchain

https://doi.org/10.1515/comp-2022-0258
received May 16, 2022; accepted September 12, 2022

Abstract: In an era of open data sharing, the scientific research field puts forward an urgent need for the value of big data. However, big data still form "data islands," which seriously affects the level of scientific research and the progress of scientific research. In this regard, this article proposes the research and realization of the big data scientific research model and key mechanism based on blockchain. This article uses the *K*-means algorithm to cluster scientific research data and reasonably utilizes the decentralization, smart contracts, and non-tampering characteristics of the blockchain to design a distributed data model based on the blockchain. This article proposes that a BIZi network is formed based on a blockchain Interplanetary File System (IPFS) and Zigzag code (blockchain, IPF Sand Zigzag code, BIZi for short) to achieve reliable data connection and through a set of data access control mechanisms and data service customization mechanism to effectively provide data requirements for scientific research. Finally, IPFS network transmission speed performance can better meet the needs of scientific research. The larger the number of file blocks, the higher the fault tolerance rate of the scheme and the better the storage efficiency. In a completely open data-sharing scenario, the fault tolerance rate of Byzantine nodes is extremely high to ensure the stability of the blockchain. The current optimal consensus algorithm fault tolerance rate reaches 49%.

Keywords: blockchain research, big data, scientific research model, key mechanism, clustering algorithm

1 Introduction

Big data refers to the amount of data involved that is too large to be captured, managed, processed, and organized into information that can help companies make more active business decisions within a reasonable period through mainstream software tools. With the rapid popularization of applications such as social networks, smart software, mobile Internet, and Internet of Things (IoT), data mining and analysis, prediction, and analysis based on advanced technologies such as machine learning have become more accurate, and the hidden value of big data has also been unearthed. With the rapid increase in the amount of modern data, all walks of life are paying increasing attention to data mining, application, and analysis, and the new era of data opening is quietly coming.

The era of big data has affected and changed the development of many industries. For example, business, scientific research, public services, and other fields have put forward a large demand for big data. However, there is still a phenomenon of "data islands" in the research of big data. A lot of data cannot be shared in common, which affects the efficiency of data used by managers or scientific researchers. Blockchain technology has unique characteristics such as decentralization and fully distributed storage. Under the conditions of combining big data, it can play a better role, provide scientific research personnel with higher-end and convenient services, and promote the development of scientific research speed and progress.

In Reyna A's vision of the IoT, traditional devices have become intelligent and autonomous. Due to technological progress, this vision is becoming a reality, but there are still some challenges to be solved, especially in the security field, such as data reliability. However, their research has not found an effective solution to the relationship between the IoT and blockchain [1]. Under the impulse of arising enormous information innovation, Wang et al. proposed to plan a major information stage, the D2D large information stage, to successfully support

* **Corresponding author: Shengwei Wen,** Department of Science and Computer, Ganzhou Teachers College, Ganzhou, 341000, Jiangxi, China, e-mail: wenshengwei3921@163.com

the remote D2D correspondence between clients, precisely give content to suppliers, and productively do offloading for administrator smart. However, there are still some problems in the sharing process of this application, which have not been solved yet [2]. Prasad et al. have developed a resource dependence model that links big data analysis with outstanding humanitarian results through case studies (qualitative) of 12 humanitarian streams. They decide the hubs in the organization, which can apply power on key NGOs in light of comparing the assets to guarantee the production of adequate enormous information. However, this survey has limited intervention measures, and the distributed network is not complete [3].

The innovations of this article are (1) establishing a blockchain-based big data scientific research model, combined with blockchain and distributed files to form a data model, storing the key and important information needed for scientific research on the blockchain, and having the characteristics of non-falsification and traceability; (2) combining theoretical analysis and empirical analysis, after establishing a big data scientific research model, and applying it to scientific research institutions to conduct empirical experiments to understand the feasibility and deficiencies of the model research place and improve.

2 Research and implementation methods of big data scientific research model and key mechanism based on blockchain

2.1 Blockchain

Blockchain is quite possibly the most progressively arising innovation in the late years. Partially, it has carried a progressive effect on finance and different ventures, and its execution possibilities are extremely wide [4,5]. According to the viewpoint of text and information structure, blockchain can be separated into two information structures: square and chain. A square can be perceived as an advanced square used to store all exchanges inside a timeframe. The chain is a chain structure that associates this information and jellies are the first successions without interference [6,7]. Blockchain can be divided into the public chain, alliance chain, and private chain.

From a technical point of view, blockchain can be regarded as a decentralized public database (or called a public general database). It is composed of some important factors [8,9], among which blockchain technology is mainly composed of five important factors: modern encryption, distributed peer-to-peer network communication technology, distributed consensus algorithm, incentive mechanism, and plan code [10,11].

From the perspective of participants, it can be divided into private blockchains, public blockchains, and consortium chains. The public chain is open, and all users can access all blockchain node information to participate in transactions. This can be completely decentralized in a practical sense. The alliance chain is composed of member nodes participating in the alliance, but this is different [12,13]. Starting from the mechanism, all consensus mechanisms will be jointly negotiated by member institutions. Private chains are mainly used for private organizations. The rules concerning the reading and writing of permits and consent forms shall be established by the authorities. Its value lies in adjusting the internal business to achieve a unified and reliable interaction [14,15].

2.2 Big data

The value of big data lies in analyzing massive amounts of data and putting forward valuable information and knowledge from it. Increasingly, companies and scientific research institutions have established their own data centers to develop big data applications [16,17]. Under normal circumstances, the analysis and processing of big data first need to aggregate the collected data and transmit it to data centers distributed in different places for processing. Data centers also often perform operations such as virtual machine migration and data off-site disaster recovery. There are a large number of terminals that need long-distance large-scale data transmission to the end.

The large-scale sharing of scientific research data not only expands the breadth and depth of the research field, but also effectively shortens the scientific research cycle, reduces the cost of data storage and management, and makes a large contribution to technological progress. Among them, data sharing standard applications for scientific research include the release and data, as well as the reuse of data from large computers, scientific instruments, and equipment. Government departments and large scientific research institutions in relevant countries have adopted relevant data-sharing policies. The central storage management will be stored in a publicly accessible third-party database so that the data of various departments and fields can be shared uniformly, and data can be downloaded.

2.3 Data clustering algorithm

The so-called clustering is the process of grouping similar things together and dividing dissimilar things into different categories, which is an essential method in data analysis. Clustering is a machine-learning technique that involves grouping data points. Given a set of data points, we can use a clustering algorithm to classify each data point into a specific group in the image.

2.3.1 *K*-means algorithm

The most classic clustering algorithm is the *K*-means algorithm, and its optimization goal formula is as follows:

$$\min J_0(C, V) = \sum_{k=1}^{K} \sum_{x \in c_k} \|x - v_k\|_2^2. \tag{1}$$

K-means obtains the solution of the local optimal equation by learning the expected value maximization algorithm. From a statistical point of view, the *K*-means tool can be regarded as a model statistical combination algorithm. In the following formula, we obtain the sample x source needed in CK from the Gaussian distribution:

$$\text{Gauss}(x; v_k, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\|x - v_k\|_2^2}{2\sigma^2}\right). \tag{2}$$

According to the known data set DB, the model parameters (C, V) are minimized:

$$\begin{aligned} &\arg\min_{(C,V)} - \sum_{k=1}^{K} \sum_{x \in c_k} \ln \text{Gauss}(x|v_k, \sigma) \\ &= \arg\min_{(C,V)} \sum_{k=1}^{K} \sum_{x \in c_k} \ln(\sqrt{2\pi}\sigma) + \sum_{k=1}^{K} \sum_{x \in c_K} \frac{\|x - v_k\|_2^2}{2\sigma^2}, \end{aligned} \tag{3}$$

$$\begin{aligned} &\arg\min_{(C,V)} - \sum_{k=1}^{K} \sum_{x \in c_k} \ln \text{Gauss}(x|v_k, \sigma) \\ &= \arg\min_{(C,V)} \sum_{k=1}^{K} \sum_{x \in c_k} \|x - v_k\|_2^2. \end{aligned} \tag{4}$$

The *K*-means algorithm is dedicated to solving the same density of model applications. Therefore, to improve the efficiency of grouping *K* machine algorithms for uneven data, we propose a new Gaussian mixture model to distinguish the differences in the number and density of groupings and then propose a new type of uneven data aggregation algorithm on this basis.

2.3.2 Non-uniform data clustering algorithm

In uneven data sets, the density of the system is usually different. Two symbol sets are inserted to explain this difference. Therefore, Sk2/wk can be used to reproduce the change of sky attribute data distribution. If substituted into the Gaussian variation of the density function (2), the probability density function of $x \in c_k$ shown by feature j is as follows:

$$p(x_j; v_{kj}, w_{kj}, \sigma_k) = \frac{1}{\sqrt{2\pi}\sigma / \sqrt{w_{kj}}} \exp\left(-\frac{(x_j - v_{kj})^2}{2\sigma^2 / w_{kj}}\right). \tag{5}$$

On this basis, based on the simple assumption that the data set feature *D* is statistically independent, a set of models is established. From the boundary distribution product of the crossover probability, only the total number of density variables of the crossover probability can be inferred. Therefore, $P(x)$ represents the probability density of any sample within CK.

$$p(x; v_k, w_k, \sigma_k) = \prod_{j=1}^{D} p(x_j, v_{kj}, w_{kj}, \sigma_k). \tag{6}$$

Then, we need to consider that the same data may contain clusters of different sizes. For this reason, the symbol αk ($k = 1,2,., K$) representing the size of the cluster is introduced to meet the constraints:

$$\forall k : \alpha_k > 0, \sum_{k=1}^{K} \alpha_k = 1. \tag{7}$$

This value is related to the number of samples included in the group and can be regarded as the weight given to each group. According to these definitions, the weighted probability function for uneven data is as follows.

$$L(\Theta) = \prod_{k=1}^{K} \prod_{x \in c_k} \alpha_k \times p(x; v_k, w_k, \sigma_k). \tag{8}$$

3 Research and implementation of big data scientific research model and key mechanism based on blockchain

3.1 Big data research model based on blockchain

The blockchain-based large information-sharing model comprises four sections including the information demander, the decentralized information-sharing stage, the information proprietor, and the information source. Information demanders allude to specialists, research foundations, organizations, or undertaking groups who need information. They can be

isolated into business applications, logical examination, public administration, and so forth, as indicated by various application fields. In this model, each subject plays a clear part in limits, and jobs can rely upon one another and change. The two players can recover information, view information quality assessment, distribute information on membership prerequisites, and so on. At last, the two players can share and collaborate with trusted, straightforward, and equivalent information authorizations on the stage.

The analysis of the three main characteristics of the blockchain model is as follows: (1) Decentralization: The update and information recording of the blockchain network are completed interactively by distributed entities, not executed by an authority. At present, the subjects of big data are widely distributed, and their relationships are equal. They all have information exchange and storage requirements and enjoy the same rights and obligations. (2) Non-tamperable: Different subjects have different requirements for data sharing. (3) Smart contracts: Big data sharing involves various data permission interactions, and the logic of the data evaluation system can ensure the automatic execution of information flow through a series of smart contracts. This state change can avoid the interference of single-point failures, malicious attacks, and other factors [18]. Figure 1 is a distributed data connection model based on blockchain.

3.2 Key mechanism of big data based on blockchain

3.2.1 Reliable data connection mechanism based on BIZi network

This article proposes that the BIZi network is formed based on blockchain Interplanetary File System (IPFS) and Zigzag code (blockchain, IPF Sand Zigzag code, BIZi) to realize reliable data connection. The IPFS is a network transfer protocol designed to create persistent and distributed storage and sharing of files, and it is a content-addressable peer-to-peer hypermedia distribution protocol.

Among them, IPFS, as the most popular distributed version file, has functions such as content addressability, highlight point hypermedia, disseminated capacity, transmission convention, document forming, and so forth, which plans to supplant the HTTP convention and assemble a more secure and more productive Internet [19]. It has great

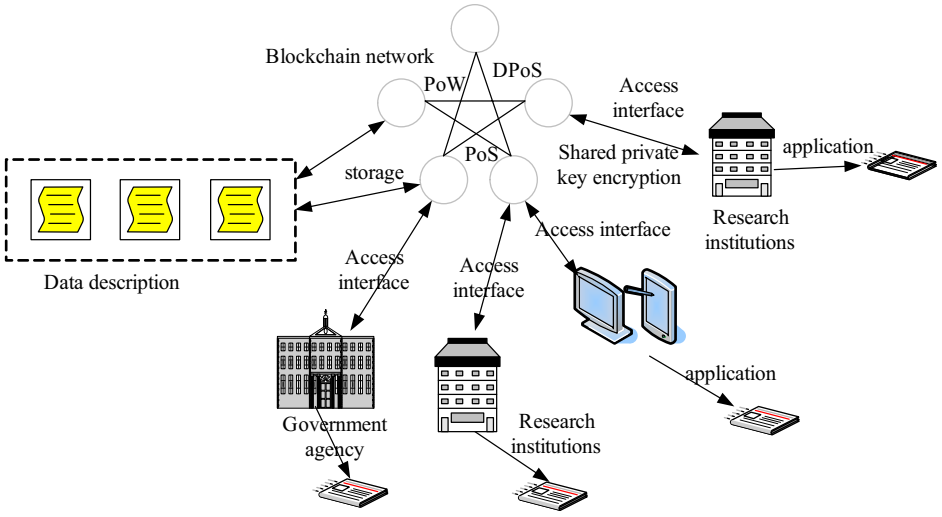

Figure 1: Blockchain-based distributed data connection model.

Table 1: BIZi technical function table

Technology	Features
Blockchain	Record the address list of each node, the hash of each routing file, and the corresponding MDS-Zigzag offset matrix
PDF	File fragment storage and access
Zigzag coding	File compression encryption and decryption restoration
BIZi	Reliable data routing and storage, greatly reducing the storage space of small files with year-on-year fault tolerance

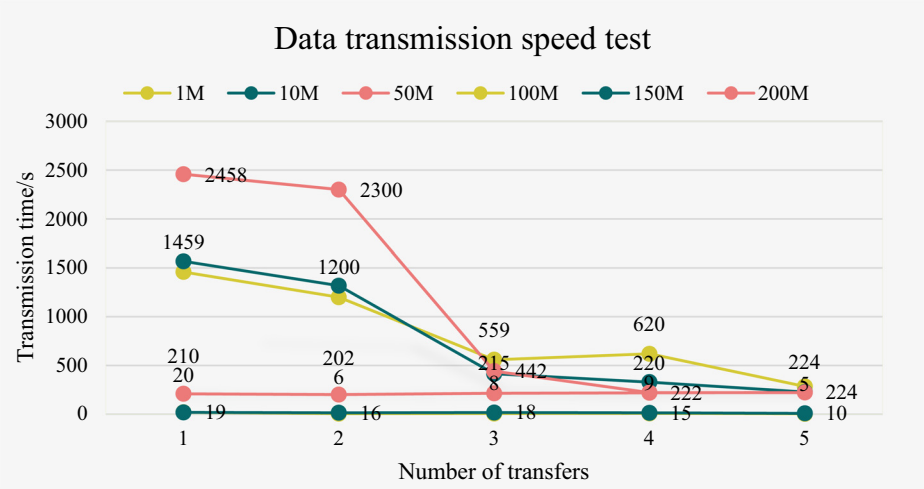

Figure 2: Data transmission speed test.

coupling with the big data connection scheme based on blockchain in this article. The specific analysis is as follows: (1) System architecture: Unlike the current mainstream distributed file system that uses a decentralized clustering method, the role of the administrator is different. IPFS uses a completely open P2P architecture to achieve a flat and open network. It is consistent with the current blockchain system and can better adapt to the distributed characteristics of big data providers. (2) Performance and security: IPFS has an elastic network link mechanism that is independent of the Internet backbone, which can greatly improve network transmission efficiency and reduce the network delay, save network bandwidth, and provide efficient and safe information flow transmission. (3) Compatibility: IPFS provides HTTP transmission methods, compatible nodes on different platforms, and visual node file management and transmission tools so that nodes can be easily deployed on a variety of terminals, which can last and effectively ensure the authenticity of files. (4) File function: IPFS not only provides block storage of large files but also provides file history version management functions. There is no file type and file size limit, so it can adapt to different types of data description forms. The function summary of the three technologies in this mechanism is shown in Table 1.

3.2.2 Blockchain-based data access control mechanism

This article presents a blockchain-based information access control component. Among them, the power-based data access control method is a typical resource management mechanism. It uses power as a special token, which can achieve fine-grained access control based on resources, convenient for permission updates, and support for permission withdrawal. Due to its strong scalability and manageability, this article combines it with the blockchain to solve the data trust problem of centralized management, and through transparent accounting to ensure that the data supply and demand parties can perform flexibly and reliably interactive.

4 Research and implementation analysis of big data scientific research model and key mechanism based on blockchain

4.1 Big data transmission speed test

Ethereum hubs joining the organization can consequently synchronize existing exchange information, take an interest in keeping up with network security and soundness, and have versatility and unwavering quality. To check the framework's simultaneousness and capacity productivity, the IPFS network transmission speed execution test results are displayed in Figure 2. The utilization of the Ethereum blockchain to record the center connection interaction can completely address the issues of information sharing [20]. The IPFS access time for records under 10 M can be settled inside the 20 s, and the transmission time for documents under 100 M is steady at 200 s, so it is achievable to utilize IPFS for non-business center record stockpiling. The experimental outcomes demonstrate the way that the model can address the issues of logical examination in execution and has been improved and culminated in different perspectives such as unwavering quality and adaptability.

Table 2: Mainstream blockchain

Blockchain type	Leading agency	Consensus mechanism	Consensus speed	Theoretical fault tolerance	Incentives
Ethereum	Community	POW + POS	15 s	49%	Token
EOS	IBM	DPOS	3 s	48%	Token
BCOS	Wanxiang Blockchain Alliance	PBFT	Second level	33%	Scenes
Neo	Neo company	PBFT	Second level	33%	Token
Fabric	Hyperledger Alliance	RAFT	Second level	0% (Assuming no Byzantine nodes)	Scenes

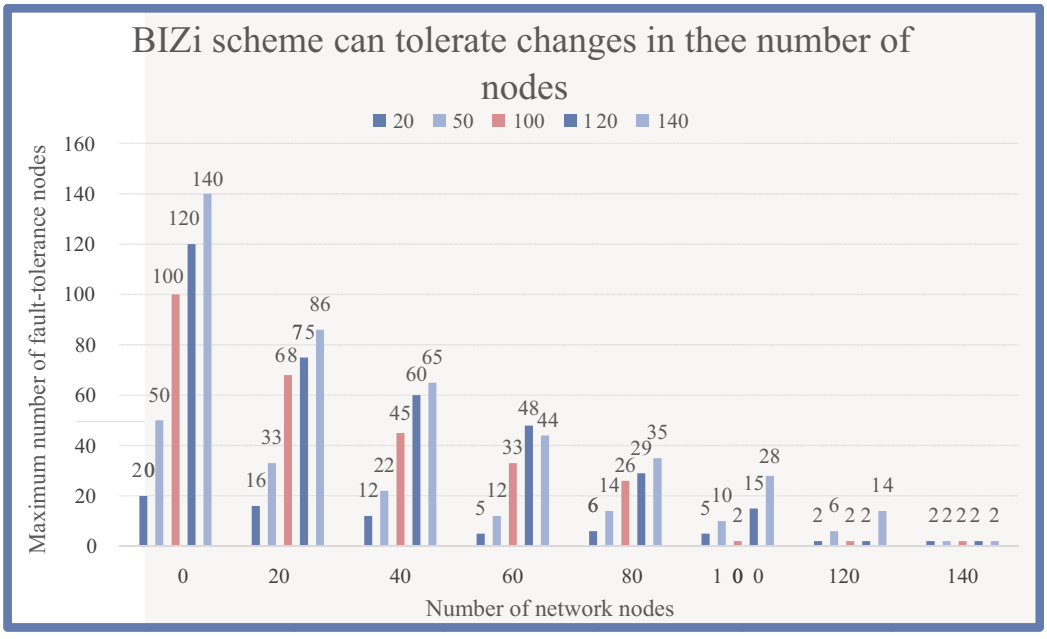

Figure 3: BIZi scheme can tolerate changes in the number of nodes.

4.2 Mainstream blockchain performance comparison

Table 2 shows that in the case of complete credibility, the blockchain platform led by the alliance is more efficient and can save costs. However, if you want to face a completely open data-sharing scenario, the fault tolerance rate of Byzantine nodes is extremely high to ensure the stability of the blockchain. The current optimal POW consensus algorithm requires no more than 51% of Byzantine nodes; that is, the fault tolerance rate is 49%. Ethereum is an open-source public blockchain platform with smart contract functions. It provides a decentralized Ethereum virtual machine to process peer-to-peer contracts through its dedicated cryptocurrency, Ethereum, with a consensus speed of 15 s.

4.3 BIZi scheme can tolerate changes in the number of nodes

Figure 3 shows that as the number of file blocks is larger, the fault tolerance rate of the scheme is higher, and the storage efficiency is better. That is to say, this solution is for small files and can provide better backup effects than large files, and as the number of nodes increases, the larger the number of file blocks that can be backed up, the longer the distance to the minimum threshold.

5 Conclusion

This article concentrates on the exploration and execution of the large information logical examination model

and a key component in the light of the blockchain. On this basis, we adopted the consensus algorithm based on the blockchain, combined with the large-scale clustering algorithm based on *K*-means, to realize the decentralization of the blockchain. Using the BIZi network, scientific data can be reliably connected, data access control, data customization, and effective data demand for scientific research. The innovation of this article is that, first of all, a big data scientific research model based on blockchain is established, which combines the blockchain and distributed files to form a data model and stores the key and important information needed for scientific research on the blockchain. And it has the characteristics of nontampering and traceability. Second, the theoretical analysis and empirical analysis are combined to establish a big data scientific research model and apply it to scientific research institutions to conduct empirical experiments to understand the feasibility and deficiencies of the model research and make improvements.

Conflict of interest: Author states no conflict of interest.

Data availability statement: Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

References

[1] A. Reyna, C. Martín, J. Chen, E. Soler, and M. Díaz, "On blockchain and its integration with IoT. Challenges and opportunities," *Future Gener. Comput. Syst.*, vol. 88, no. NOV, pp. 173–190, 2018.

[2] X. Wang, Y. Zhang, V. C. Leung, N. Guizani, and T. Jiang, "D2D big data: Content deliveries over wireless device-to-device sharing in large scale mobile networks," *IEEE Wirel. Commun.*, vol. 25, no. 1, pp. 32–38, 2018.

[3] S. Prasad, R. Zakaria, and N. Altay, "Big data in humanitarian supply chain networks: a resource dependence perspective," *Ann. Oper. Res.*, vol. 270, no. 1, pp. 383–413, 2018.

[4] E. Mengelkamp, B. Notheisen, C. Beer, D. Dauer, and C. Weinhardt, "A blockchain-based smart grid: towards sustainable local energy markets," *Comput. Sci. Res. Dev.*, vol. 33, no. 1–2, pp. 207–214, 2018.

[5] M. Möser, K. Soska, E. Heilman, K. Lee, H. Heffan, S. Srivastava, et al., "An empirical analysis of traceability in the monero blockchain," *Proc. Privacy Enhancing Technol.*, vol. 2018, no. 3, pp. 143–163, 2018.

[6] M. H. Miraz and M. Ali, "Applications of blockchain technology beyond cryptocurrency," *Ann. Emerg. Technol. Comput.*, vol. 2, no. 1, pp. 1–6, 2018.

[7] H. Jang and J. Lee, "An empirical study on modeling and prediction of bitcoin prices with bayesian neural networks based on blockchain information," *IEEE Access*, vol. 6, pp. 5427–5437, 2018.

[8] G. Liang, S. R. Weller, F. Luo, J. Zhao, and Z. Y. Dong, "Distributed blockchain-based data protection framework for modern power systems against cyber attacks," *IEEE Trans. Smart Grid.*, vol. 10, no. 3, pp. 3162–3173, 2018.

[9] F. Gao, L. Zhu, M. Shen, K. Sharif, Z. Wan, and K. Ren, "A blockchain-based privacy-preserving payment mechanism for vehicle-to-grid networks," *IEEE Netw.*, vol. 32, no. 6, pp. 184–192, 2018.

[10] V. Sharma, I. You, F. Palmieri, D. N. Jayakody, and J. Li, "Secure and energy-efficient handover in fog networks using blockchain-based DMM," *IEEE Commun. Mag.*, vol. 56, no. 5, pp. 22–31, 2018.

[11] L. Li, J. Liu, L. Cheng, S. Qiu, W. Wang, X. Zhang, et al., "CreditCoin: A privacy-preserving blockchain-based incentive announcement network for communications of smart vehicles," *IEEE Trans. Intell. Transport. Syst.*, vol. 19, no. 99, pp. 2204–2220, 2018.

[12] Y. Chen and Y. Chi, "Harnessing structures in big data via guaranteed low-rank matrix estimation," *IEEE Signal. Process. Mag.*, vol. 35, no. 4, pp. 14–31, 2018.

[13] A. R. Al-Ali, I. A. Zualkernan, M. Rashid, R. Gupta, and M. AliKarar, "A smart home energy management system using IoT and big data analytics approach," *IEEE Trans. Consum. Electron.*, vol. 63, no. 4, pp. 426–434, 2018.

[14] G. Ke, D. Tao, J. F. Qiao, and W. Lin, "Learning a no-reference quality assessment model of enhanced images with big data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 4, pp. 1301–1313, 2018.

[15] N. Zhang, P. Yang, J. Ren, D. Chen, L. Yu, and X. Shen, "Synergy of big data and 5G wireless networks: Opportunities, approaches, and challenges," *IEEE Wirel. Commun.*, vol. 25, no. 1, pp. 12–18, 2018.

[16] P. Antonetti and S. Maklan, "Identity bias in negative word of mouth following irresponsible corporate behavior: A research model and moderating effects," *J. Bus. Ethics*, vol. 149, no. 4, pp. 1–19, 2018.

[17] T. Ritter and C. Lett, "The wider implications of business-model research," *Long. Range Plan.*, vol. 51, no. 1, pp. 1–8, 2018.

[18] H. H. Emira, "Authenticating IoT devices issues based on blockchain," *J. Cybersecur. Inf. Manag.*, vol. 1, no. 2, pp. 35–40, 2020.

[19] V. Mani, P. Manickam, Y. Alotaibi, S. Alghamdi, and O. I. Khalaf, "Hyperledger healthchain: Patient-centric IPFS-based storage of health records," *Electronics*, vol. 10, p. 3003, 2021.

[20] O. I. Khalaf and G. M. Abdulsahib, "Optimized dynamic storage of data (ODSD) in IoT based on blockchain for wireless sensor networks," *Peer-to-Peer Netw. Appl.*, vol. 14, no. 5, pp. 2858–2873, 2021, doi: 10.1007/s12083-021-01115-4.