FDC: A Secure Federated Deep Learning Mechanism for Data Collaborations in the Internet of Things

Bo Yin, Hao Yin, Yulei Wu, *Senior Member, IEEE*, and Zexun Jiang

***Abstract*—With the explosive network data due to the advanced development of the Internet of Things (IoT), the demand for multiparty computation is increasing. In addition, with the advent of future digital society, data have been gradually evolving into an effective virtual asset for sharing and usage. With the nature of the sensitivity, massiveness, fragmentation, and security of multiparty data computation in the IoT environment, we propose a secure data collaboration framework (FDC) based on federated deep-learning technology. The proposed framework can realize the secure collaboration of multiparty data computation on the premise that the data do not need to be transmitted out of their private data center. This framework is empowered by public data center, private data center, and the blockchain technology. The private data center is responsible for data governance, data registration, and data management. The public data center is used for multiparty secure computation. The blockchain paradigm is responsible for ensuring secure data usage and transmissions. A real IoT scenario is used to validate the effectiveness of the proposed framework.**

***Index Terms*—Data collaboration, data privacy, deep learning, federated learning, Internet of Things (IoT).**

I. Introduction

With the emergence of new information technologies, such as Internet of Things (IoT) [1] and mobile payment, human society, physical world, and cyberspace are further converged and integrated. The era of future digital society in which human beings and characters are deeply integrated is coming. As a link between the physical world and cyberspace, data have penetrated into every aspect of people's lives and work and have evolved into a commercial capital and an important economic input. The information contained in the data reflects the indiscrimination of human labor and machine power and can provide productivity and knowledge guidance for improving the efficiency of production and enhancing the quality of human lives. The data contain big value, which makes it an effective asset in the future digital society [2]. Data to the digital society are like the fuel to the industrial revolution. Full exploitation and utilization of data can promote many new products and services and can also enhance the security, stability, and development of the national and social economy. For example, the data generated from personal lives and work constitute the personal assets of data owners. These data include basic personal information, personal income and expenditure, personal health, personal education, and other forms of raw data and need to be owned or controlled by individuals. Digital assets are being accumulated in cyberspace through advanced information technologies, which will cause qualitative changes from quantitative changes and will bring revolutionary impact on the business model of many industries [3].

However, the current Internet service model is application centric, and the premise for users to enjoy the convenience of Internet applications is to share their data, e.g., preferences, with application providers. This results in the situation that "data follow the application, and my data are owned by others," which leads to the monopoly of data by the industry giants who manage and control the mainstream applications [4]. Data abuse may lead to illegal profits, uncontrolled privacy and limited innovation, weakening governance functions, and causing other serious problems. Moreover, the deficiencies in the trust mechanism, data management, and the process security of existing Internet systems, which were mainly designed for information dissemination, can further lead to the issues of unclear data ownership, unclear sources, lack of supervision of transactions, and inability of assessment and verification. These problems seriously hinder the development of IoT and the evolutionary development of the industries and the coming digital society.

Data are the essence of the digital society which connect everything, for example, people, machines, and material things on earth, and the data have important value [5]. It is an inevitable trend for data to be an asset in the development of digital society. If the data become an asset, it must be controllable, measurable, and monetizable. However, the characteristics of data replication and the current development of Internet applications make the data difficult to control, measure, and monetize. Besides, it is hard to define data

Manuscript received October 10, 2019; revised December 17, 2019; accepted December 30, 2019. Date of publication January 15, 2020; date of current version July 10, 2020. This work was supported in part by the National Key Research and Development Program under Grant 2016YFB1000102; and in part by the National Natural Science Foundation of China under Grant 61972222 and Grant 61672318. *(Corresponding author: Hao Yin.)*

Bo Yin is with the School of Computer Science, Beijing Information Science and Technology University, Beijing 100101, China (e-mail: raul_yinbo@163.com).

Hao Yin and Zexun Jiang are with the Research Institute of Information Technology, Tsinghua University, Beijing 100084, China (e-mail: h-yin@mail.tsinghua.edu.cn; jiangzx14@mails.tsinghua.edu.cn).

Yulei Wu is with the College of Engineering, Mathematics and Physical Science, University of Exeter, Exeter EX4 4QF, U.K. (e-mail: y.l.wu@exeter.ac.uk).

Digital Object Identifier 10.1109/JIOT.2020.2966778

ownership. The management of data value is, therefore, a challenging problem in the distribution of multiparty interests in IoT, hindering the fast development of IoT.

With the rapid development of information technologies, e.g., artificial intelligence, deep learning, big data, IoT, and blockchain, the era of digital society has become a reality [6], [7]. In this article, we develop a secure data collaboration mechanism based on the federated learning paradigm, named FDC to deal with secure collaboration of massive data for multiple parties in an IoT environment. Specifically, the framework consists of a private data center, a public data center, and a blockchain system. A private data center is responsible for IoT data collection, storage, management, and registration. The data computation is performed in the private data center to ensure that data does not need to be transmitted outside the private cloud. A public data center with a heterogeneous and distributed data storage system is designed to support scattered data fragments with cryptography protection, improving the intrinsic safety of the whole system. It is worth noting that the data calculated in the public data center are different from those in the private data center. A blockchain-based data access authorization system is devised to support flexible and efficient access control for the entities with different incentives based on a token scheme. In addition, a unified and flat authorization logic system is developed to update the mapping of data owners and users to improve the efficiency of data authorization. A distributed consensus mechanism based on the blockchain and distributed ledger is proposed to achieve reliable records of behaviors, further realizing multiparty incentive and privacy protection. A case study is carried out to validate the feasibility of the proposed FDC mechanism.

The remainder of this article is organized as follows. Section II presents the related challenges. Related work is summarized in Section III. The overview of the proposed system is described in Section IV. A case study is given in Section V to validate the effectiveness and feasibility of the proposed mechanism. Finally, Section VI concludes this article.

II. Research Challenges

IoT has been developed for many years but still it has many outstanding problems, such as different data standards, difficulty of defining data ownership [8], lack of credible incentives for data sharing [9], and difficulty of tracing data behavior [10]. These problems can result in data fragmentation in the network. This further causes the issue of “Everyone has data, but everyone cannot control their data and cannot own the value of their data.” The contradiction and dilemma of “people who produce the data are lack of data” hinder the data sharing and trading.

A. Data Ownership Ambiguity

As a future core asset, the definition of property rights of data is vague under current technical conditions [11], [12]. The data governance has not been fully established. For example, the underlying rule of the current industry is “who collects data, who owns data.” Internet companies collect a large volume of personal data of users and make profits by processing, using, and disseminating these data. Due to the lack of data ownership confirmation mechanisms in the network, the infringement of users’ data property rights, privacy rights, and income rights usually happens [13]. The related ethical and legal issues become increasingly prominent if the data and algorithms involve the sharing process. Data disclosure scandals have made the public feel that their privacy has been violated. Several large companies have been resisted and criticized by the public and have been investigated and punished by the government. Therefore, we urgently need a mechanism to keep data in the private data centers owned by data producers, while allowing data to participate in computation.

With the increasingly prominent problem of data ownership and data privacy, several studies have focused on privacy protection in the data-exchange networks [14], [15], and the government has begun to promulgate laws and regulations for data sovereignty and privacy protection [16], restraining the accountability for data operations from the perspective of supervision. For example, in May 2018, the EU promulgated the General Data Protection Regulation [17], which specifies the behavioral norms that Internet enterprises should follow when collecting, storing, computing, and disseminating user data and gives users more data autonomy and control rights. These regulations provide a reference for the formulation of laws and regulations in the field of data protection and extend the mechanism of data ownership protection from the technical issues of cyberspace to the rules of human society. With the rapid progress of data capitalization, more formal data ownership protection documents will be issued.

B. High Risk of Data Exchange

Data are often stored in the servers of different application providers, and data storage entities with different interests manage the data under their own control. Due to the lack of effective confirmation mechanisms for data ownership in cyberspace, it is difficult to guarantee effective traceability and economic incentives for data publishing, transferring, and usage. Data owners may face problems such as loss of ownership, loss of revenue, and legal liability after data abuse [18]. Data consumers may also face problems, such as buyer fraud and data quality, which lead to the unwillingness of both data owners and data consumers to take risks to sell or purchase the data voluntarily. This seriously affects the circulation, sharing, and usage of data.

Data decentralization is the data stored in the storage space of different departments. Data access authorization management rules and the system architectures of different departments are often incompatible. Cross-domain authorization of data needs to be negotiated offline according to the policies and characteristics of different departments, which leads to the complexity of data access authorization processes.

C. Scattered Data Management

Each application provider collects user data through its client and stores and manages user data through the background servers [19]. The same user may interact with multiple applications at the same time, resulting in its user data collected by multiple applications. For example, the social communication software collects the users' social chat records, the search engines collect the users' Web browsing habits, the online banking and third-party payment platforms collect the users' financial transaction information, the browser and input methods collect the users' operational habits, and the e-commerce platforms collect the users' commodity transaction and payment information [20]. Many of these applications collect and store the same user's data, resulting in the isolation of user data within different application providers. Due to the different business characteristics and business demands of different application providers, data access has become a challenging problem. The phenomenon of data islands leads to the decentralization and fragmentation of data management. It is difficult for the users' data to form a unified view in cyberspace, and it is not conducive to the collection, management, analysis, and usage of multidimensional personal data.

D. Insecurity in Data Collaborative Computing

With the rapid development of Internet technologies, emerging network applications have the need of cross-domain collaborative computing, which requires the fusion and collaboration of a large number of network data. In this process, the data belonging to different stakeholders and with different ownerships will be aggregated into network applications. All data privacy information may be exposed to application providers, which inevitably leads to the risk of losing the control of data privacy [21].

For the sake of security, most of the data of IoT devices are stored in private data centers, and the storage cost of expensive construction and management are carried out by private data centers [22]. In many situations, people use a third-party application, and the data are generated by the application provider. Because the data are generated, stored, and managed by different application providers, the ownership of data is out of control. There is no effective redundancy guarantee mechanism to ensure data safety and privacy. Data control depends entirely on the management strategy of third-party service providers and the third-party data damage caused by technical weaknesses, operational errors, policy risks, and so on may lead to irreparable losses for data.

In the present circumstances, data sharing behavior is recorded in the server of application providers, lacking of multiconsensus witness. The integrity, consistency, and duration of data behavior records are limited by their own internal resources and control scope. The cross-domain behavior traceability is, therefore, difficult to obtain the evidence.

III. Related Work

The current Internet system provides a certain degree of trust protection for data interaction in the application layer. This makes the data sovereignty controlled by the application provider. In other words, it results in the situation of data sovereignty out of control, where "data follow the application and my data are owned by others." Throughout the construction of information highway in the past decades, although the network has been widely interconnected and interoperable, the lack of built-in security incentives and credible mechanisms makes it difficult to guarantee the unification of security for data sharing and circulation and reliable distribution of data rights and interests. This results in the abuse and illegal reuse of data. Therefore, how to design and improve the future network architecture to overcome these issues has become a research hotspot in both academia and industry.

In terms of guaranteeing data ownership and secure data sharing, the famous Internet infrastructure, domain name service (DNS), publicizes the availability of corresponding data through the binding relationship between domain name and IP address, so as to realize the publicity and circulation control of network data. However, DNS relies on its hosting agency. Because the existing architecture is centralized management, the data tampering is difficult to be found. This results in a high risk of data out of control. The distributed hash table (DHT) technique for indexing data in a decentralized way improves the efficiency of data sharing [23], but lack of pairwise. The traceability mechanism of data circulation processes is difficult to solve the problem that digital objects can be copied and used at will. This leads to the difficulty of realizing the unification of data sharing, and rights and interests recognition and distribution. There is a high risk of data abuse. In this case, the rights and interests of data providers are often not adequately protected.

The common problem of the traditional Internet technology lies in the tight coupling between application and data. To use DNS and other services, users need to submit their data to the application providers according to the specifications of application protocols, which inevitably leads to user data occupied by the application providers. To cope with this dilemma, decoupling data and applications and creating a new data-centric Internet application model has gradually attracted the attention of academia and industry [24]. For example, the solid project led by Tim Berners-Lee, the father of WWW, is based on the idea that the data control is handed over to users. The MIT Connection Science Laboratory has put forward the idea of future information infrastructure, including a set of mechanisms, such as digital identity, trusted distributed computing, data storage, and application execution separation. It is suggested that applications should be implemented in a specific regulatory environment rather than in the application providers' servers. MIT Media Lab, a media lab in China, has made a deep research on the new storage mechanism to protect data ownership and privacy. Through its openPDS project, it has created a new data service mode that can hide user data privacy information by using a "question-response" model. Domestic scientific research circles have also put forward the concept of "personal data bank" to protect data ownership and circulation expediency and have constructed a data authorization access and orderly circulation system to provide management and value-added services of large data assets while safeguarding user data sovereignty, so as to enhance the users data revenue.

The introduction of the blockchain technology has brought a new hope for the data security management. Its distributed account ledger cannot be tampered with, and the transaction records can be trusted and consensus. It provides natural security for data security management, ownership confirmation, and rights and interests circulation [25]. Wang *et al.* [26] proposed a blockchain-based solution for securing the data-exchange process. It can be achieved without relying on any centralized application providers, which greatly promotes the emergence of the Internet of Values and even the future digital society [27]. Reference [28] systematically expounds the data structure, data model, and data storage scheme of the blockchain data. This provides significant guidance for the safe storage paradigm of network data and facilitates the users to design corresponding data management schemes according to data ownership guarantee and safe circulation requirements. Reference [29] proposed LinkShare. It tracks the process of storage, usage, and circulation of user data based on the distributed characteristics of blockchain. Cha *et al.* [5] designed a low-power and high-efficiency data access authorization and privacy protection mechanism based on the blockchain cryptography mechanism for the IoT scenario with resource-constrained devices. In the aspect of medical data management with high privacy requirement, MIT Media Lab developed the functions of registration, parsing, and sharing of medical data based on an intelligent contract platform. The authorized access of users' medical data should be validated and authorized by the corresponding intelligent contract. In this way, the unified management and rights acquisition of medical data in different hospitals can be realized. To solve reliable registration and definition of data ownership, the research team led by Robert Khan, one of the inventors of TCP/IP, proposed a digital object architecture [30], a decentralized system called Namecoin [31]. The blockstack [32] provided the mechanism and method guarantee for the decentralized data asset ownership confirmation and publicity. Moreover, the blockchain technology is expected to build a reliable identification and confirmation system for data objects. This system is difficult to tamper with and has clear ownership and credible process of circulation and transaction due to its untouchable and traceable characteristics of data on the chain. For example, in the UAV data collection and communication scenarios, the hash fingerprint information of data collected by UAV is linked as data identification by using the above characteristics of blockchain technology. The data exchange between UAVs is realized by sharing data identification flows. The data flow records of the whole communication process can be divided into blockchains. The complete traceability in the distributed account ledger realizes the data integrity and auditability of the whole process [33].

In addition, the blockchain system is based on the consensus mechanism to ensure that all participants form a consistent view of transaction behavior in digital space [34]. The recognized consensus record has the advantages of nontampering and traceability, which makes reliable traceability and auditing of various network application behaviors, supported by intelligent contracts, become a reality [28], [35], [36]. It makes the system possess the characteristics of data security and trustworthy credit self-establishment [37]. It provides technical supports for trustworthy data flow and value transfer between untrustworthy entities in virtual cyberspace (such as encrypted currency systems [38], [39], which realize the trustworthy value transfer between anonymous participants). At the same time, [39] also provided convenience for the establishment of trustworthy data flow and value transfer between anonymous participants. A secure and credible future digital society lays the technical foundation [40]. Based on the advantages of blockchains in trusted data flow and value transfer, VMware, a leading global computing infrastructure and mobile commerce enterprise, proposes to integrate blockchain technology into a cloud computing platform to support trusted data transmission services [41]. However, its data storage and organization still depend on the centralized cloud storage service addresses, and data sovereignty protection relies on cloud service providers rather than the users, which makes it difficult to deal with the risk of data abuse effectively. Castro *et al.* [42] applied the blockchain technology to the negotiation and validation of interconnection strategies among Internet autonomous systems. The proposed scheme supports collaboration and economic incentives based on the advantages of blockchain. To improve the efficiency of interdomain routing between autonomous domains, the reliable interconnection at the Internet routing level has been realized. The MeDShare scheme [43] was based on the level of healthcare applications, using the blockchain technology in the untrustworthy network environments to establish the trust among multiple parties in medical scenarios. The paper [44] and our previous work [45] all mentioned the idea of building the future value Internet and proposed to build the future information infrastructure which can guarantee the security and reliability of data storage, calculation, and circulation.

IV. Proposed System

The proposed FDC mechanism aims to solve the problems of data ownership confirmation, efficient authorization, secure and confidential storage, secure sharing and efficient management, cross-domain cooperative secure computing, and reliable traceability and audit of data behaviors.

A. System Architecture

Fig. 1 is the architecture of the proposed FDC mechanism. It consists of three layers: 1) data layer; 2) control layer; and 3) application layer. In the *data layer*, the private data center stores the data uploaded by the data publisher through the IoT devices such as Apple watches. It processes and governs the data to a certain data unit to facilitate the data sharing. In the application layer, the data registration platform registers the data unit uploaded by the data publisher from private data centers, to protect data ownership. The data management platform that manages the data, possessing ownership relation with data publishers, and guarantees the data publishers' ownership of the data. In the *control layer*, the public data center is responsible for securing multiparty computation. In the circulation of data collaboration, the blockchain records the behavior of the data usage to ensure that the behavior of the data is authentic.

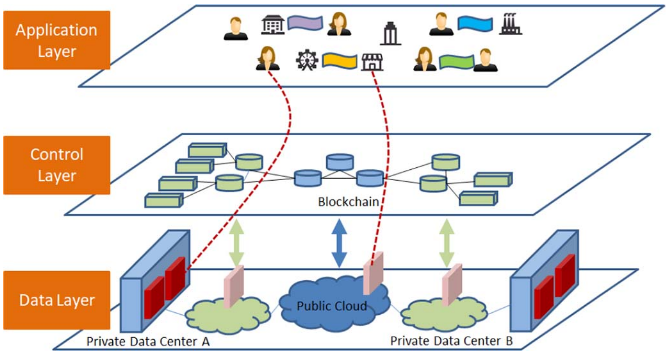


Fig. 1. System architecture of the proposed FDC mechanism.

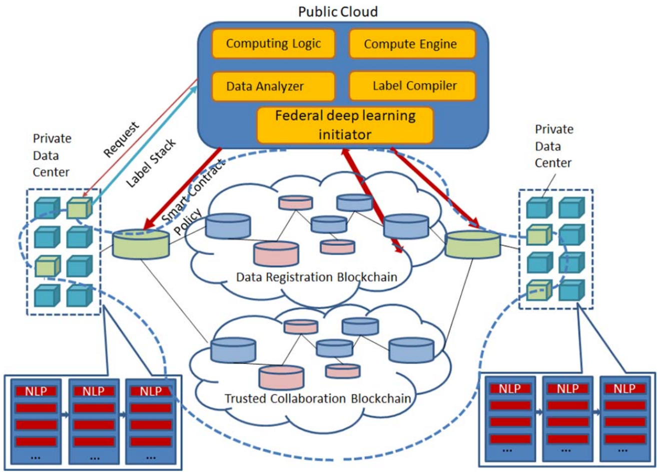


Fig. 2. Components of the proposed FDC mechanism.

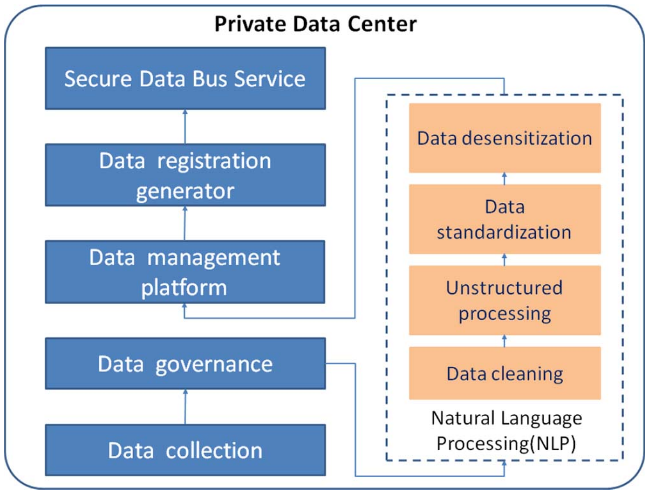


Fig. 3. Main functions of a private data center.

When the target data is used by a data user, it is processed in the proposed system. The processing result is then returned to the data user. This can thereby realize the separation of data ownership and usage rights and ensure that the data circulation process is secure and controllable. The federated learning of data for multiparty computing is completed in the public data center.[1] Federated learning realizes data calculation without using the data itself and ensures the security of user data, i.e., the data still stays in the private data center.[2]

The functions of the private data center, the public data center, and the collaboration blockchain are described in detail below.

The collaboration between the components mentioned above is shown in Fig. 2. The natural language processing (NLP) platform is configured to process data uploaded by a data publisher to generate a data unit. The data registration blockchain is configured to register, according to the data unit, the data uploaded by the private data center to identify the data ownership. This blockchain is also configured to manage the data which has an ownership relationship with the data publisher. Each data publisher corresponds to the respective private data center. The trusted collaboration blockchain is configured to record collaboration and trading behavior of data collaboration, such as federated deep learning of data. The public cloud is configured to process the federated deep learning of different private data centers and return the processing result to the federated deep learning initiator.

B. Private Data Center

The main functions of private data centers are shown in Fig. 3. The private data center is to manage the data and generate data units that are easy to share and confirm ownership. The main functions contain data collection, data governance, NLP, data registration generator, and secure data bus service. *Data collection* is responsible for data collection. *Data governance* is using the NLP method to realize the data cleaning, unstructured processing, data standardization, and data desensitization. After data governance, the *data registration generator* generates the preprocessed data to specific data units in order to facilitate data sharing and circulation. The remaining function is *secure data bus service* which realizes the data interaction with other systems.

During the data registration generator, the data format in the data processing phase is difficult to be unified, and the data processing efficiency is low because of the diversity of data types. Therefore, we propose to use the unified content label (UCL) as the carrier of the data, which is registered as a data tag on this registration platform. The method proposed in this article can effectively improve the data processing efficiency and achieve better data services at a lower cost. The unified content tag is a string of fixed-length key-value pairs, including the name, profile, hash, timestamp, and address of the data. The data user can obtain the summary of the data from the unified content tag, and then determines whether it is necessary to further obtain the complete set of data from the data address. An example of the format of a unified content tag is shown in Fig. 4.

The data registration generator is an innovation point of the proposed secure data collaboration system. It ensures that each data in the system can have a unique digital identity, which facilitates the tracking of the collaboration and circulation behavior of the data assets. As shown in Fig. 4, the UCL

[1] The terms of public data center and public cloud are used interchangeably in this article.

[2] The terms of private data center and private cloud are used interchangeably in this article.

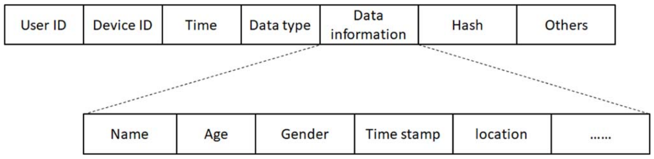


Fig. 4. Example of UCL to describe personal data.

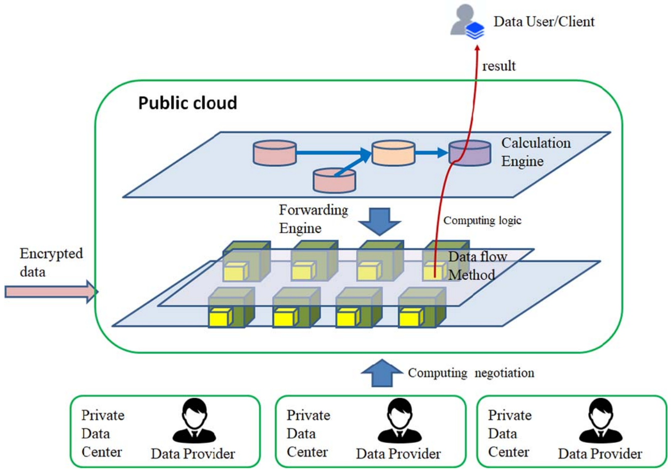


Fig. 5. Secure computing mechanism and its workflow in public cloud.

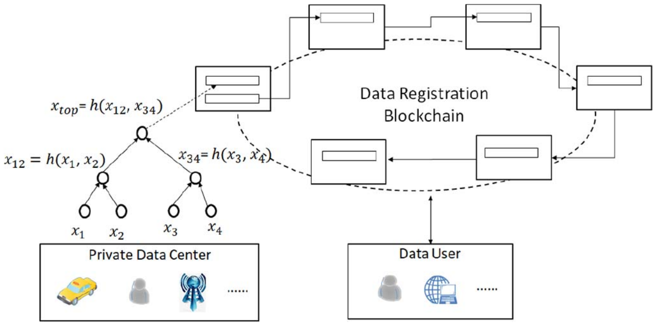


Fig. 6. Function and node structure of data registration blockchain.

is the main output of data registration generator which can be used to realize many functions, such as transaction account identification, registration and authentication, citation analysis, version history, signature verification, etc. The citation acquisition, historical change, exchange and transaction, derivative development, and interconnection and integration calculation of data in the system are all based on unified naming standards. The UCL distributed identity authentication function realizes the same data validation service as the Internet-domain name resolution system and, then, solves the problems of inefficient data object case processing and the difficulty of content-driven routing in address-driven Internet.

C. Public Data Center

The current data circulation and sharing scheme rely on the traditional Internet infrastructure with given routing paradigms to achieve data flow between terminals. However, when the traditional routing is used as the underlying information infrastructure, the traditional data circulation scheme has a bad performance. The perception of the data itself lacks the necessary supervision and traceability for data circulation behaviors, and the security performance is poor. Moreover, the traditional data circulation behavior is usually implemented by point-to-point transmission. The traditional routing method is not flexible, and it involves the value conversion problem [*] associated with the corresponding data flow.

This article proposes a secure computing mechanism that includes a data flow method, a calculation engine, and a forwarding engine as shown in Fig. 5. First, the *calculation engine* is used to obtain the data transmission request from a client, which carries the target data description. The public cloud responds to the data transfer request, finds the target data to be calculated in the private cloud, and returns the result of the target data sharing behaviors to the client. After the *forwarding engine* receives the target data and the target circulation behavior definition, the calculation engine processes the target data according to the target circulation behavior definition. We present the target circulation behavior definition in the calculation engine. When transmitting the target data, we use the forwarding engine to perform the target data circulation behavior according to the present target circulation behavior definition. The *dataflow method* refers to data organization and processing methods, which complete data calculation in multiple private data centers. The proposed method can effectively solve the problem of poor security and flexibility of the existing data circulation and sharing schemes.

D. Blockchain

To deal with the difficulty in guaranteeing data consistency and the loss of data in the process of data collaboration and sharing, we combine the blockchain technology to design a blockchain-based data registration mechanism and trusted collaboration mechanism. Through the data registration blockchain, the ownership of the data can be clarified and the consistency of the data can be guaranteed.

To solve the data consistency problem by our proposed data registration blockchain, the first step is to synchronize the root hash value of the data among nodes using blockchain. The user can obtain the root hash value of the data through each node, and then synchronize them through peer-to-peer communication. The data provider obtains each hash value on the hash path of the required file and, then, the user determines whether the root hash value obtained from the node is consistent by verifying the consistency of the hash value of obtained data and the metadata. The registration process is shown in Fig. 6. First, the registration process manages the data in the private cloud to generate UCL and, then, it hashes the UCL and broadcasts it to the registration blockchain for registration. The hash value is written in the ledger of the blockchain. The entire registration process is then complete.

The data registration blockchain is the core function of organizing, managing, and recording data information. It supports the registration and publicity of complete data description information. It also supports the interaction between blockchain, security collaboration computing, and other components to achieve cross-validation of data usage behaviors and provides historical records of data transactions and data version changes. The functions, such as dynamic recording, facilitate fine-grained statistical analysis of the data flow behaviors of the system.

The trusted collaboration blockchain includes a data identification module, a pass-through management module, a smart contract module, and a consensus module. The *data identification module* is used for registration before data uploading, and its function is to assist the user to complete standardized description of the data. The *pass-through management module* is used to implement the pass-based authority verification for the user's access to the data. The *smart contract module* is proposed to support the compilation and operation of the smart contract on the license chain, and support more flexible implementation of data collaboration strategy. The *consensus module* in the blockchain includes consensus algorithms.

The workflow of trusted collaboration blockchain is described as follows. The data user initiates a data request to the trusted collaboration platform by using the clients and provides a token list for the authority verification. The trusted collaboration blockchain interacts with the registration blockchain and returns the uploading authority token and the user data registration description of the client. The security collaboration and circulation functions provided in the private data center introduce the technology of trusted collaboration blockchain to store and supervise the data behavior. It mainly solves the problem of auditing the participants' behavior and supervising the compliance in the public cloud.

According to the digitalization characteristics of data, the trusted collaboration blockchain provides data behavior certificate and life-cycle management functions and designs the life-cycle traceability mechanism of relevant data. Every data upload request, data access request, data access authorization, data storage, data sharing, data collaborative computing, and other actions in the system are recorded in this blockchain, forming a complete record of the whole process of data behavior and ensuring the untouchability of data behavior records through untouchable distributed accounts.

Compared with traditional centralized audit, the introduction of blockchain technology can enhance the tamper-proof, consistency, traceability, reliability, and authority control of audit records as shown in Fig. 6. The security of data is guaranteed by the trusted collaboration blockchain. Once the data is linked, the implementation mechanism of blockchain can ensure the consistency, tamper-proof, repudiation-proof, and reliability of audit data in the chain. In addition, blockchains are used to store all the audit data. For data access on the chain, it can only be accessed by specific roles after signature verification. At the same time, the system integrates the authorization module based on the permission pass, which improves the strength of data access control. For the authenticity of the data under the chain, the authenticity of the identity can be ensured by the signature mechanism of the data initiator. The cross-validation of the data initiated by different participants can identify whether there is forged data. In addition, the source code audit of the system components can verify whether the system components have been tampered with or embedded in the malicious code.

The trusted collaboration blockchain has the function of detection and warning of violations. Moreover, compared with the traditional audit system, it can enhance the tamper-proof, consistency, traceability, reliability, and authority control of audit records.

E. Client

The client provides the operation for data collaboration. It integrates digital transaction interfaces, such as distributed data set catalog, algorithm libraries, and algorithm models to provide users with an easy-to-use operation interface.

The client provides network private key and security hardware management, and only interacts with the proposed system through the security hardware device (U shield). The basic process of the client works as follows: the data user inquires the registration blockchain and initiates a data collaboration request through the client. The client signs the data access or sends the data collaboration request through the security hardware device. The signed request is then transmitted by the client to the trusted collaboration blockchain which returns the request result to the client after interacting with the registration blockchain. It can be seen that the client is only a visual interface system for data users and does not involve any core business process.

V. Case Study: Federated Learning on Wearable Sensor Data

To evaluate our proposed mechanism, we use Libra [46] to realize the trusted collaboration blockchain and the data registration blockchain. Then, we implement a federated deep learning algorithm in the public cloud to realize multiparty secure computing.

The public cloud submits a smart contract with data number, data type they want, and the token quantity expected to pay. And then, the private clouds sign the data amount they can provide. After that, we get into the processing of the contract, where first, we move the total tokens from the public cloud address to the contract address. Then, the contract gets the data amount transferred from every private cloud. Finally, we calculate how many tokens every private cloud can get, pay for them, and return the left tokens to the public cloud.

To achieve this, we use a private blockchain based on the Libra protocol which allows a set of replicas referred to as validators from different authorities to jointly maintain a database of programmable resources. These resources are owned by different user accounts authenticated by public key cryptography and adhere to customize the rules specified by the developers of these resources. The validators process the transactions and interact with each other to reach consensus on the state of the database. Some of the private cloud and public cloud are deployed as the validator nodes to serve the blockchain.

The workflow of blockchain is shown in Fig. 7. The smart contract interface for initiating federated deep learning is shown in Fig. 8. In Fig. 7, people who want to finish a federated deep learning by using multiparty data, start a smart contract in the public cloud which contains data number, data type, and the total number of token. Then, the public cloud

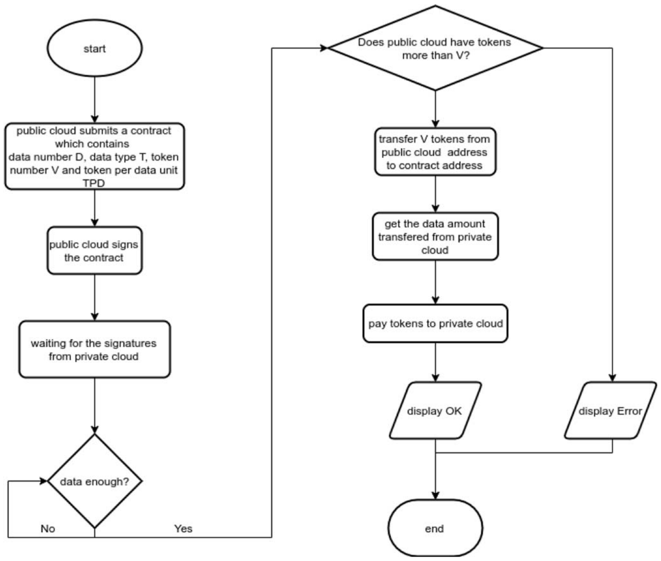
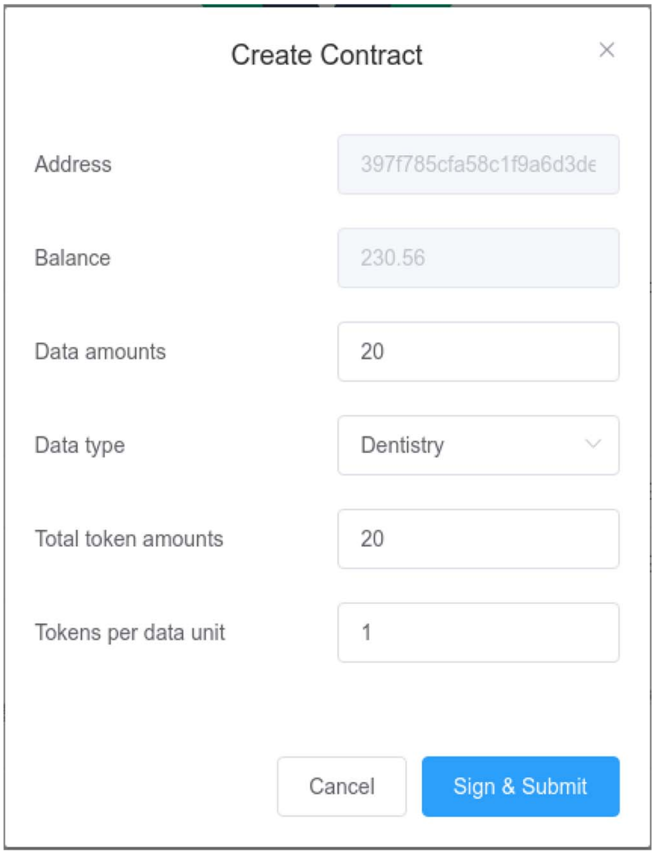

start
public cloud submits a contract which contains data number D, data type T, token number V and token per data unit TPD
public cloud signs the contract
waiting for the signatures from private cloud
data enough?
No
Yes
Does public cloud have tokens more than V?
transfer V tokens from public cloud address to contract address
get the data amount transfered from private cloud
pay tokens to private cloud
display OK
display Error
end

Fig. 7. Workflow of blockchains.

signs the contract and waits for the signatures from the private cloud. The trusted registration collaboration cloud prepares the data from multiparty private clouds and estimates whether the data satisfied the public cloud demand. If the data are enough and the public cloud has enough token to pay these data, the blockchain transfers the tokens from public cloud to private cloud using the contract address. The process of trust collaboration by blockchain is then completed. In Fig. 8, the address is the private cloud which provides data and gets token from data users. The data amount is the requested data by the public cloud. The balance is the remaining tokens that can be distributed. Data amounts and data type are the description about the requested data. Total token amounts and tokens per data unit are the description about the token that can be distributed. Smart contracts allow trusted transactions and agreements to be carried out among disparate, anonymous parties without the need for a central authority, a legal system, or an external enforcement mechanism. After the process of the smart contract, transactions are recorded on the blockchain which cannot be tampered with and can ensure authenticity and safety. In our case, the smart contract is developed using the language called "move," which is a new programming language developed to provide a safe and programmable foundation for the Libra blockchain. We abstract it again to form the templates in the medical scene to make it more friendly and easier to use. People in the hospital can use it conveniently without the knowledge of the blockchain and smart contracts.

One of the key components of the proposed FDC mechanism is to adopt the federated learning, which can preserve data privacy and optimize system cost. To demonstrate the process of federated learning, we propose a practical scenario of children physical activity (PA) tracked by wearable devices. With the development of movement sensor and devices, there are many PA tracking applications. For example, smart watches, such as iWatch, can track exercise process, and smartphone can track everyday walking steps of people. Most of these applications need to manually designate the PA types to accurately model and track the activities. The basic function of PA tracking is to acquire user physical features, such as calorie consumption for weight control, sleep status tracking, etc.


Create Contract

Address 397f785cfa58c1f9a6d3de

Balance 230.56

Data amounts 20

Data type Dentistry

Total token amounts 20

Tokens per data unit 1

Fig. 8. Smart contract creation process.

PA tracking can be useful for children education. For example, if we can obtain a clear and accurate PA status of a child, the schools or kindergartens can arrange exercise and meal plan accordingly, which can be helpful for obesity control. However, the long-term PA tracking for children is challenging. The major challenge is to establish the connections between sensor data and PA. Small children cannot precisely log their PA through mobile applications. To overcome this challenge, we collect the data from several kindergarten in Beijing. A customized wearable IoT devices are designed to log movement accelerated velocity. In addition, the PAs of children are logged by teachers.

To establish the connections, deep learning technology is utilized. However, the collected data may contain privacy information of children, so that the schools are not willing to share the data to public. Federated learning can allow the data to be used without sharing or revealing privacy information. The experimental results are used to verify the effectiveness of the proposed architecture. In the future, large-scale applications need to be verified on a larger scale of data.

A. Data Set

Table I shows the key parameters of the collected raw data. The data set consists of two parts, the movement data and PA logs. This table shows the basic characters about our collected data.

TABLE I
DATA SET PARAMETERS

Sampling frequency	4000Hz
Start Time	2019-09-04
Duration	41 Hours
Subject Number	30
Range of acceleration	-8g 8g

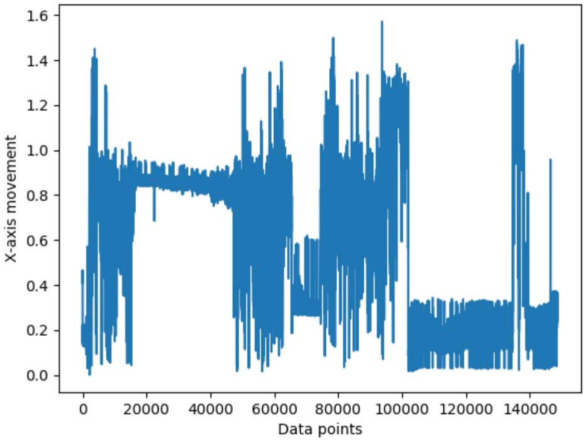


Fig. 9. *x*-axis movement example of children for front and back movement.

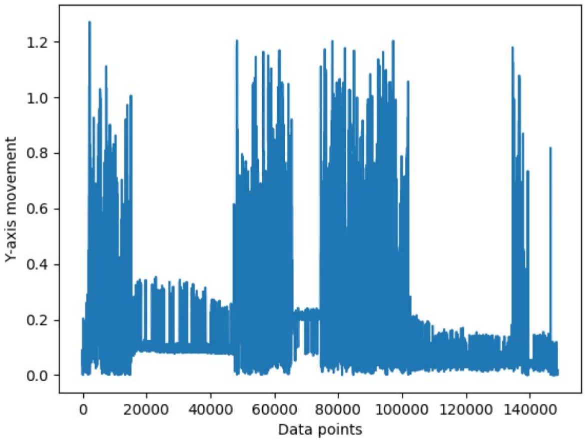


Fig. 10. *y*-axis movement example of children for left and right movement.

The movement data are collected by wearable devices which measure and log the accelerated velocity from 3-D at a high frequency. Three-dimensional representation of children's three moving directions include front and back, left and right, and up and down. The accelerated velocity is between 8g and −8g. To compress the data and features for deep learning, the raw data is combined in seconds by the following equation:

$$A_c(s) = \sum_{a_i \in s} a_i \Delta t \tag{1}$$

where a_i is a data point, and s is a time period of one second. Δt is the time interval between data points. An example of combined movement data is shown in Figs. 9–11. From these figures, it indicates that some of the patterns can be distinctive.

PA logs are logged by the teachers in the schools, as shown in Table II.

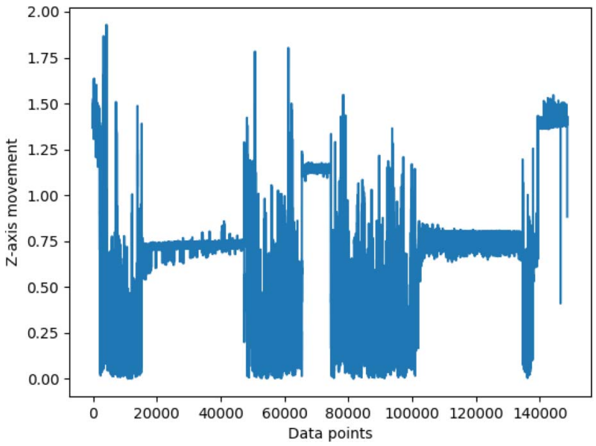


Fig. 11. *z*-axis movement example of children for up and down movement.

TABLE II
PA LOG TABLE

Time	Activity
8:00 8:20	Breakfast
8:20 9:45	Teaching
9:45 10:00	Indoor activities
10:10 11:00	Outdoor activities
11:00 11:30	Indoor activities
11:30 12:00	Lunch
12:00 14:00	Afternoon nap
14:30 16:10	Outdoor activities
16:10 16:50	Dinner

The PAs of teaching are not specific, so the activity is not included in the data set. Based on the log, we classify all PAs into four categories.

1) *Eating:* Breakfast, lunch, and dinner.
2) *Sleeping:* Afternoon nap.
3) Indoor activities.
4) Outdoor activities.

The goal of the federated learning is to use the movement data to obtain the PA information.

B. Problem Formation

With the movement data $\vec{A}_c$, the problem is to establish the prediction function f for the above four types of PAs.

$$c = f\left(\vec{A}_c\right) \tag{2}$$

where c represents a type of PA.

C. Metrics

To evaluate the training and performance of federated learning, the following metrics are utilized. Because accuracy and loss are usually used to express the effectiveness of deep learning.

1) *Accuracy* is the ratio of correctly classified data points.

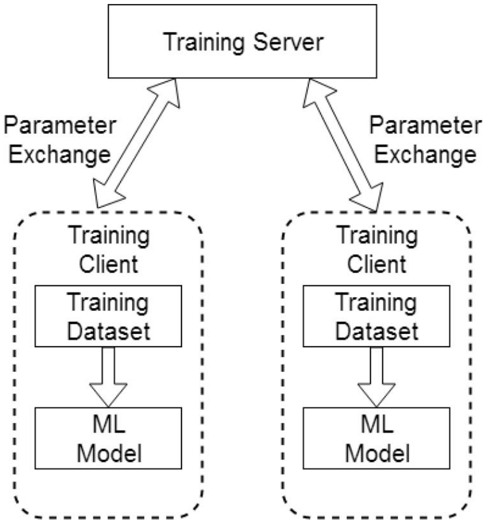


Fig. 12. Federated learning paradigm.

2) *Loss* function is the categorical cross entropy (CCE), expressed by

$$\mathrm{CCE} = -\sum_{i}^{C} t_i \log(s_i) \tag{3}$$

where t_i and s_i are the ground truth and model-prediction score for each class i.

D. Federated Learning

The goal of federated learning is to train a centralized model while training data remains distributed. The federated learning framework is shown in Fig. 12. While remaining competitive performance compared with centralized training, the federated learning can reduce the cost of transferring data through network and protect privacy information, since only model parameters are exchanged [47].

Tensorflow federated learning [48] is utilized to simulate the process. It uses some models and data to simulate and verify the federated learning algorithms embedded in it. These metrics, such as accuracy and loss, are calculated automatically by Tensorflow. The federated learning process can be divided into the following steps.

Preprocess: In preprocessing, a minute consists of 60 data points (seconds). Since each data point (in the unit of second) consists of three accelerated speed on x-axis, y-axis, and z-axis, the data point (in the unit of minute) is a 180-length vector.

Model: To demonstrate the case, we use a general neural network model. The model consists of four layers.

1) Input layer can receive the feature vector with 180 floats.
2) Dense layer (fully connected) with 1024 cells.
3) Dense layer (fully connected) with 128 cells.
4) Dense layer with softmax as output layer.

A scenario of five distributed training clients is simulated, which represents that there are five independent schools conducting the research, as the client shown in Fig. 12. The federated learning server utilizes simple parameter average function as the aggregation process.

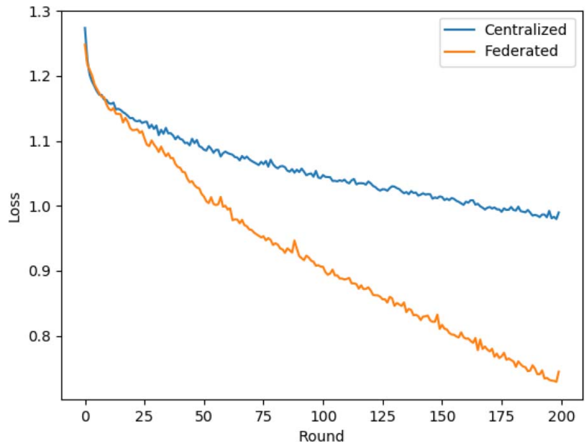


Fig. 13. Loss during the training of the model.

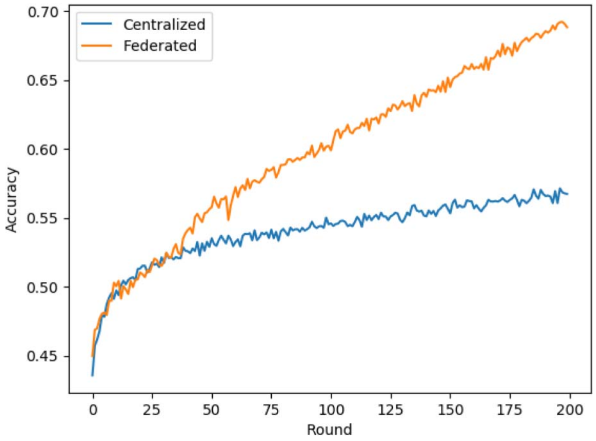


Fig. 14. Accuracy during the training of the model.

TABLE III
PERFORMANCE OF THE MODEL BY TEST DATA SET

Method	Loss	Accuracy
Federated	1.111	0.5406
Centralized	1.559	0.5176

E. Training and Evaluation

There are 35 different test subjects (Children). Thirty of them are used as the training sets, and five are used for testing. Each client trains the neural model with six subjects. All methods are trained in 200 rounds.

Federated learning is compared with traditional centralized learning with the same neural network model. Figs. 13 and 14 show the training process. It shows that the centralized learning converges faster than federated learning. It is due to the overly simple parameter aggregation process. In most cases, the centralized computing method has better computing efficiency than the distributed computing method. However, through fine-grained adjustment, the efficiency of distributed computing can approach the centralized computing method, while ensuring the data security sharing. As can be seen from Figs. 13 and 14, the loss during the training of federated deep learning is lower than the centralized method, and the accuracy is higher.

As for evaluation by test data sets, Table III shows the overall results. It indicates that federated and centralized learning perform similarly on the test data sets. The federated learning performs slightly better than centralized learning. Since only a basic general model is utilized, the performance has improvement potential. The results show that federated learning can achieve similar results as the centralized learning, which proves the practical value of our framework.

VI. Conclusion

This article has presented a secure data collaboration framework based on blockchain and federated learning for the IoT environment. Aiming at the sensitivity and intensity of IoT data, we have proposed the framework to realize the collaboration for the computation of multiparty data on the premise that data does not need to be transmitted outside of a private cloud. To solve the fragmentation of IoT data, we have adopted a private data center to govern data into UCL. To address privacy and security issues of IoT data, we have developed a blockchain-based mechanism to record multiparty interactions. To ensure large-scale multiparty secure collaborations of IoT data, we have adopted the federate learning to tackle it. Finally, we have validated the feasibility of the proposed framework in a real-world IoT scenario. Through federated learning for children's movement data, multiparty secure computation can be realized.

References

[1] J. Liu, B. Krishnamachari, S. Zhou, and Z. Niu, "DeepNap: Data-driven base station sleeping operations through deep reinforcement learning," *IEEE Internet Things J.*, vol. 5, no. 6, pp. 4273–4282, Dec. 2018.

[2] O. B. Sezer, E. Dogdu, and A. M. Ozbayoglu, "Context-aware computing, learning, and big data in Internet of Things: A survey," *IEEE Internet Things J.*, vol. 5, no. 1, pp. 1–27, Feb. 2018.

[3] H. Zhu, Y. Cao, X. Wei, W. Wang, T. Jiang, and S. Jin, "Caching transient data for Internet of Things: A deep reinforcement learning approach," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 2074–2083, Apr. 2019.

[4] P. Li, Z. Chen, L. T. Yang, Q. Zhang, and M. J. Deen, "Deep convolutional computation model for feature learning on big data in Internet of Things," *IEEE Trans. Ind. Informat.*, vol. 14, no. 2, pp. 790–798, Feb. 2018.

[5] S.-C. Cha, J.-F. Chen, C. Su, and K.-H. Yeh, "A blockchain connected gateway for BLE-based devices in the Internet of Things," *IEEE Access*, vol. 6, pp. 24639–24649, 2018.

[6] X. Zhang, L. Yao, S. Zhang, S. S. Kanhere, M. Sheng, and Y. Liu, "Internet of Things meets brain–computer interface: A unified deep learning framework for enabling human-thing cognitive interactivity," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 2084–2092, Apr. 2019.

[7] J. Mills, J. Hu, and G. Min, "Communication-efficient federated learning for wireless edge intelligence in IoT," *IEEE Internet Things J.*, to be published.

[8] A. Mashhadi, F. Kawsar, and U. G. Acer, "Human data interaction in IoT: The ownership aspect," in *Proc. Internet Things*, 2014, pp. 159–162.

[9] B. E. Bierer, M. Crosas, and H. H. Pierce, "Data authorship as an incentive to data sharing," *New England J. Med.*, vol. 377, no. 4, p. 402, 2017.

[10] B. Yu, J. Wright, S. Nepal, L. Zhu, J. Liu, and R. Ranjan, "IoTChain: Establishing trust in the Internet of Things ecosystem using blockchain," *IEEE Cloud Comput.*, vol. 5, no. 4, pp. 12–23, Jul./Aug. 2018.

[11] T. Fisher, *The Data Asset: How Smart Companies Govern Their Data for Business Success*. Hoboken, NJ, USA: Wiley, 2009.

[12] J. Dugast and T. Foucault, "Data abundance and asset price informativeness," *J. Financial Econ.* vol. 130, no. 2, pp. 367–391, Mar. 2017.

[13] M. Hall *et al.*, "Rethinking abstractions for big data: Why, where, how, and what," *J. Comput. Sci.*, vol. 1, pp. 1–8, Mar. 2013.

[14] Z. Yao, J. Ge, Y. Wu, and L. Jian, "A privacy preserved and credible network protocol," *J. Parallel Distrib. Comput.*, vol. 132, pp. 150–159, Oct. 2019.

[15] Y. Ma, Y. Wu, J. Li, and J. Ge, "APCN: A scalable architecture for balancing accountability and privacy in large-scale content-based networks," *Inf. Sci.*, vol. 9, pp. 1–22, Jan. 2019.

[16] Y. Amar, H. Haddadi, and R. Mortier, "Privacy-aware infrastructure for managing personal data," in *Proc. ACM SIGCOMM Conf.*, 2016, pp. 571–572.

[17] P. Voigt and A. V. D. Bussche, *The EU General Data Protection Regulation (GDPR): A Practical Guide*. Cham, Switzerland: Springer Int., 2017.

[18] X. Yue, H. Wang, D. Jin, M. Li, and W. Jiang, "Healthcare data gateways: Found healthcare intelligence on blockchain with novel privacy risk control," *J. Med. Syst.*, vol. 40, no. 10, p. 218, 2016.

[19] S. D. Liang, "Smart and fast data processing for deep learning in Internet of Things: Less is more," *IEEE Internet Things J.*, vol. 6, no. 4, pp. 5981–5989, Aug. 2019.

[20] W. Raghupathi and V. Raghupathi, "Big data analytics in healthcare: Promise and potential," *Health Inf. Sci. Syst.*, vol. 2, no. 1, p. 3, 2014.

[21] T. S. Raghu and H. Chen, *Cyberinfrastructure for Homeland Security: Advances in Information Sharing, Data Mining, and Collaboration Systems*, Elsevier Sci., 2007.

[22] R. Ying, H. W. Li, and L. Wang, "A new data collaboration service based on cloud computing security," in *Proc. IOP Conf. Mater. Sci. Eng.*, 2017, pp. 1–7.

[23] I. Stoica *et al.*, "Chord: A scalable peer-to-peer lookup protocol for Internet applications," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 31, no. 4, pp. 149–160, 2001.

[24] T. Chajed *et al.*, "Amber: Decoupling user data from Web applications," in *Proc. USENIX Conf. Hot Topics Oper. Syst.*, 2015, p. 19.

[25] A. Dorri, M. Steger, S. S. Kanhere, and R. Jurdak, "BlockChain: A distributed solution to automotive security and privacy," *IEEE Commun. Mag.*, vol. 55, no. 12, pp. 119–125, Dec. 2017.

[26] K. Wang, J. Dong, Y. Wang, and H. Yin, "Securing data with blockchain and AI," *IEEE Access*, vol. 7, pp. 77981–77989, 2019.

[27] G. Zyskind, O. Nathan, and A. S. Pentland, "Decentralizing privacy: Using blockchain to protect personal data," in *Proc. IEEE Security Privacy Workshops*, San Jose, CA, USA, 2015, pp. 180–184.

[28] K. Christidis and M. Devetsikiotis, "BlockChains and smart contracts for the Internet of Things," *IEEE Access*, vol. 4, pp. 2292–2303, 2016.

[29] A. Banerjee and K. P. Joshi, "Link before you share: Managing privacy policies through blockchain," in *Proc. IEEE Int. Conf. Big Data*, 2017, pp. 4438–4447.

[30] P. Svärd, "Enterprise content management and the records continuum model as strategies for long-term preservation of digital information," *Rec. Manag. J.*, vol. 23, no. 3, pp. 159–176, 2013.

[31] T.-H. Chang and D. Svetinovic, "Data analysis of digital currency networks: Namecoin case study," in *Proc. Int. Conf. Eng. Complex Comput. Syst.*, 2017, pp. 122–125.

[32] 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, pp. 173–190, Nov. 2018.

[33] M. S. Ali, K. Dolui, and F. Antonelli, "IoT data privacy via blockchains and IPFS," in *Proc. Int. Conf. Internet Things*, 2017, pp. 1–7.

[34] D. Yermack, "Corporate governance and blockchains," *Rev. Finance*, vol. 21, no. 1, pp. 7–31, 2017.

[35] M. Conoscenti, A. Vetrò, and J. C. D. Martin, "Blockchain for the Internet of Things: A systematic literature review," in *Proc. IEEE/ACS 13th Int. Conf. Comput. Syst. Appl.*, 2017, pp. 1–6.

[36] P. K. Sharma, S. Singh, Y.-S. Jeong, and J.-H. Park, "DistBlockNet: A distributed blockchains-based secure SDN architecture for IoT networks," *IEEE Commun. Mag.*, vol. 55, no. 9, pp. 78–85, Sep. 2017.

[37] L. Li *et al.*, "CreditCoin: A privacy-preserving blockchain-based incentive announcement network for communications of smart vehicles," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 7, pp. 2204–2220, Jul. 2018.

[38] A. Narayanan, J. Bonneau, E. Felten, A. Miller, and S. Goldfeder, *Bitcoin and Cryptocurrency Technologies: A Comprehensive Introduction*. Princeton, NJ, USA: Princeton Univ. Press, 2016.

[39] J. Bonneau, A. Miller, J. Clark, A. Narayanan, J. A. Kroll, and E. W. Felten, "SoK: Research perspectives and challenges for bitcoin and cryptocurrencies," in *Proc. IEEE Symp. Security Privacy*, 2015, pp. 104–121.

[40] E. Anceaume, A. Guellier, R. Ludinard, and B. Sericola, "Sycomore: A permissionless distributed ledger that self-adapts to transactions demand," in *Proc. IEEE 17th Int. Symp. Netw. Comput. Appl. (NCA)*, Cambridge, MA, USA, 2018, pp. 1–8.

[41] V. Sharma, I. You, F. Palmieri, D. N. K. 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, May 2018.

[42] I. Castro, A. Panda, B. Raghavan, S. Shenker, and S. Gorinsky, "Route bazaar: Automatic interdomain contract negotiation," in *Proc. USENIX Conf. Hot Topics Oper. Syst.*, 2015, p. 9.

[43] X. Qi, E. B. Sifah, K. O. Asamoah, J. Gao, X. Du, and M. Guizani, "MeDShare: Trust-less medical data sharing among cloud service providers via blockchain," *IEEE Access*, vol. 5, pp. 14757–14767, 2017.

[44] B. Scott, J. Loonam, and V. Kumar, "Exploring the rise of blockchain technology: Towards distributed collaborative organizations," *Strategic Change*, vol. 26, no. 5, pp. 423–428, 2017.

[45] H. Yin, D. Guo, K. Wang, Z. Jiang, Y. Lyu, and J. Xing, "Hyperconnected network: A decentralized trusted computing and networking paradigm," *IEEE Netw.*, vol. 32, no. 1, pp. 112–117, Jan./Feb. 2018.

[46] *An Introduction to Libra*. Accessed: Jul. 7, 2019. [Online]. Available: https://libra.org/en-US/white-paper

[47] Y. Fu, H. Wang, K. Xu, H. Mi, and Y. Wang, "Mixup based privacy preserving mixed collaboration learning," in *Proc. IEEE Int. Conf. Service Oriented Syst. Eng. (SOSE)*, 2019, pp. 275–280.

[48] *TensorFlow Federated Learning*. [Online]. Available: https://tensorflow.google.cn/federated/federated_learning,2020-1-7

Bo Yin received the Ph.D. degree from the Beijing University of Posts and Telecommunications, Beijing, China, in 2013.

She works with Beijing Information Science and Technology University, Beijing. From January 2018 to January 2020, she worked as a postdoctoral student with RIIT, Tsinghua University, Beijing. Her research interests include cloud computing, blockchain, wireless networks, big data, and software-defined networking.

Hao Yin received the B.S., M.E., and Ph.D. degrees in electrical engineering from the Huazhong University of Science and Technology, Wuhan, China, in 1996, 1999, and 2002, respectively.

He is a Professor with the Research Institute of Information Technology, Tsinghua University, Beijing, China. His research interests span broad aspects of multimedia communication and computer networks.

Prof. Yin was elected as the New Century Excellent Talent of the Chinese Ministry of Education in 2009, and won the Chinese National Science Foundation for Excellent Young Scholars in 2012.

Yulei Wu (Senior Member, IEEE) received the B.Sc. degree (First Class Hons.) in computer science and the Ph.D. degree in computing and mathematics from the University of Bradford, Bradford, U.K., in 2006 and 2010, respectively.

He is a Senior Lecturer with the Department of Computer Science, University of Exeter, Exeter, U.K. His research has been supported by the Engineering and Physical Sciences Research Council of United Kingdom, the National Natural Science Foundation of China, and University's Innovation Platform and Industry. His expertise is on networking and his main research interests include autonomous networks, intelligent networking technologies, network slicing and softwarization, SDN/NFV, green networking, wireless networks, network security and privacy, and analytical modeling and performance optimization.

Dr. Wu contributes to major conferences on networking in various roles including the Steering Committee Chair, the General Chair, the Program Chair, and the Technical Program Committee Member. He is an editor of the IEEE TRANSACTIONS ON NETWORK AND SERVICE MANAGEMENT, *Computer Networks* (Elsevier), and IEEE ACCESS. He is a Fellow of Higher Education Academy.

Zexun Jiang received the B.S. degree in computer science from Tsinghua University, Beijing, China, in 2014, where he is currently pursuing the Ph.D. degree with the Department of Computer Science.

His research interests include network measurement and information retrieval.