Privacy-preserving and Secure Industrial Big Data Analytics: A Survey and the Research Framework

Linbin Liu, June Li, Jianming Lv, *Member, IEEE*, Juan Wang, *Member, IEEE*, Siyu Zhao and Qiuyu Lu

***Abstract*—The development of the Industrial Internet will generate a large amount of valuable data, known as industrial big data (IBD). By mining and utilizing IBD, enterprises can improve production efficiency, reduce costs and risks, optimize management processes, and innovate services and business models. However, industrial big data comes from various institutions in all walks of life and has features such as multi-source, heterogeneity, and multi-modality. And data sharing and trading (DS&T) occur in the Industrial Internet environment without mutual trust. These characteristics pose new challenges to analytics methods and privacy and security protection technologies. Therefore, this paper aims to provide references for privacy-preserving and secure industrial big data analytics (IBDA) from three perspectives: research framework, platform architecture, and key technologies. Firstly, we review the current state of research on theories and technologies related to IBDA. Then, we reveal three challenges to secure and efficient IBDA. We take the analytics and utilization of IBD as systematic engineering, propose the research framework for privacy-preserving and secure IBDA, and point out the specific content to be studied. Further, we design the architecture of the IBDA platform with the idea of layering, including a function model, security architecture, and system architecture. Finally, detailed research proposals and potential technologies for IBD analytics and utilization are presented from three aspects: data fusion and analytics, data privacy and security protection, and blockchain.**

***Index Terms*—Industrial big data (IBD), data analytics, federated learning (FL), data sharing and trading (DS&T), privacy and security, blockchain.**

I. Introduction

IN the era of Industry 4.0, the development of the Industrial Internet has become a common choice for most industrial powers to cope with future opportunities and challenges [1]. The governments of the United States, China, Japan, and Germany regard the construction of the Industrial Internet as a national strategy to develop the real economy and improve their competitive advantages by deepening the application of digital technology and promoting digital transformation. Companies such as Microsoft, Honeywell, Siemens, Alibaba, and Baidu are speeding up the construction of Industrial Internet platforms and promoting service innovation. In 2021, the added value of the global Industrial Internet industry reached $3.73 trillion [2]. Through continuous integration with emerging technologies such as big data, cloud computing, edge computing, wireless sensing, artificial intelligence, and blockchain, the Industrial Internet has realized the interconnection of various elements such as devices, products, and services. The Industrial Internet will bring innovative application modes in intelligent manufacturing, healthcare, transportation, energy, *et al*.

Industrial big data (IBD) is a term for relevant data sets in the industrial field and is a core asset of the Industrial Internet. It originates from different sectors of multiple industrial enterprises in multiple industries and is generated in all processes of industrial production. IBD is characterized by "3V" of volume, variety, and velocity and "3M" of multi-source, multi-dimension, and multi-noise [3]. *Data analytics* utilizes methods such as statistics, machine learning, and pattern recognition to uncover hidden values from massive amounts of data. This can support the elaboration and intelligent development of enterprise production, management, research, service, security, and other activities. IBD value mining can help enterprises optimize their businesses, improve quality and efficiency, and promote innovation and transformation. Sharing IBD among enterprises is valuable for the development of the industry ecosystem and its social and economic benefits. At the national level, IBD integration and analytics are significant for energy conservation, emission reduction, and protecting critical infrastructure from advanced persistent attacks [4], [5].

The realization of IBD value is based on the assumption that there will be widespread data sharing and trading (DS&T). *Data sharing* is the open usage of data among data owners within a certain scope in order to collaborate in mining data value. *Data trading* refers to the exchange of data between providers and consumers using currency or monetary equivalents, where data is regarded as a commodity. According to a survey by the Industrial Internet Industry Alliance (AII), 96% of industrial enterprises have data circulation scenarios, but 73.4% of them are concerned that the data they provide will be used for purposes other than those specified in the contract [6]. IBD is owned by various businesses and may contain commercial secrets or user private information, creating serious privacy leakage issues when shared or traded. The 2022 Cost of Data Breaches Report by IBM Security and Ponemon Institute reveals that the average cost of data breaches suffered by critical infrastructure companies such as energy, transportation, communications, and healthcare is $4.82 million [7]. About 20% of these breaches are due to the

Manuscript created June, 2023. This work was supported by the National Science Foundation of China under Grant 51977155. (Corresponding author: June Li)

Linbin Liu, June Li, Juan Wang, Siyu Zhao and Qiuyu Lu are with the Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University, Wuhan 430072, China(e-mail: linbinl@whu.edu.cn; jeli@whu.edu.cn; jwang@whu.edu.cn; 570384505@qq.com; 2018282110251@whu.edu.cn)

Jianming Lv is with the Institute of Computing Technology Chinese Academy of Sciences, South China University of Technology, Guangzhou 510641, China(e-mail: jmlv@scut.edu.cn)

compromise of business partners, i.e., supply chain attacks. Additionally, there are still many security issues to be solved in data access, transmission, storage, and utilization in the open and complex environment of the Industrial Internet.

In this paper, we reviewed big data fusion and analytics methods, data privacy and security protection methods, and blockchain technology supporting industrial big data analytics (IBDA). We also proposed a research framework for IBDA that considers privacy and security protection, and provided research ideas or technical direction references for IBDA research and platform construction.

A. Related Work

Scholars have conducted reviews on the relevant aspects of big data analytics technology and big data privacy and security.

The latest developments in big data analytics tools, methods, models, and applications, as well as research challenges, are reviewed in [8, 9]. Some academics have focused on various aspects of big data analytics in industrial scenarios. For example, Jagatheesaperumal *et al.* [10], Khan *et al.* [11] and Javaid *et al.* [12] paid special attention to key applications, technologies, and challenges. Li *et al.* [13] introduced the architecture of the Industrial Internet and discussed key enabling technologies such as big data, while Bonnard *et al.* [14] and Santos *et al.* [15] presented a big data analytics architecture for industry 4.0, respectively. Additionally, several review articles investigated the application status and challenges of IBDA technologies in different industries, including manufacturing [16–19], energy [19–23], transportation [24], construction [25, 26], healthcare [27] and logistics [28].

The privacy and security issues of big data were explored in [29, 30], which also introduced the latest security technologies and methods. In [31], the most advanced privacy-preserving big data analytic techniques were evaluated from four dimensions: data utility, robustness, complexity, and efficiency. In [32], various security and privacy solutions for big data analytics in cloud environments were reviewed from three perspectives: secure access control, secure data storage, privacy-preserving learning. A comprehensive survey on blockchain and big data was provided in [33], which summarized various blockchain services for big data, including secure data collection, data storage, data analytics and data privacy protection. Several review articles specifically covered the privacy challenges, current solutions and cutting-edge technologies of big data in agriculture [34] and healthcare [35], [36].

The aforementioned literature offers a comprehensive review of recent advancements in big data analytics techniques and methods for protecting security and privacy. It also presents an overview of the applications and challenges encountered in various industrial contexts. However, these surveys merely summarize relevant technologies and approaches, without providing specific research content to be examined from the broader perspective of data analytics and privacy protection. While several papers address open issues and future research directions in the field of IBDA, they do not propose concrete solutions. Furthermore, our findings indicate that no existing literature provides a framework considering security and privacy for an IBDA platform. Therefore, the current body of literature is insufficient to serve as a reference or guide for research into privacy-preserving and secure IBDA and platform construction.

B. Contributions

In this work, we begin by conducting a review of IBDA and security-related technologies. Subsequently, we propose a research framework for IBDA that incorporates considerations for privacy and security protection. Finally, we outline the technical approaches for each component of the framework. Our main contributions are as follows:

1) We explore the current state of theories and technologies that are related to the secure and efficient IBDA. This includes theories on big data fusion and analytics, methods for data privacy and security protection, and blockchain technologies for IBDA.
2) We propose a research framework for privacy-preserving and secure IBDA. We highlight the research content from four perspectives: platform construction, big data analytics theory, data privacy and security protection technology, and blockchain technology supporting IBDA. The framework is based on the understanding that the analytics and utilization of IBD are systematic processes that require consideration of privacy and security protection from the design of infrastructure, the integration of multiple security technologies, and coordination with big data analytics technology.
3) We suggest a function model, security architecture, and system architecture for the IBDA platform. These architectures are designed using a layered approach, taking into account the platform's functional requirements as well as the privacy and security demands at each stage of IBD collection, storage, analytics, sharing, and trading. These architectures can serve as a blueprint for constructing the IBDA platform.
4) We point out the key technologies that support secure and efficient IBDA and suggest research proposals and potential technologies. IBDA methods are discussed from three perspectives: data joint characterization, multi-modal fusion analytics, and distributed elastic computing. Privacy and security protection methods for IBD are discussed from three perspectives: data sensitive attribute identification, secure federated learning, and access control. Blockchain technologies that support secure and efficient IBD analytics and utilization are discussed from three perspectives: block structure, consensus algorithms, and smart contracts.

C. Paper Organization

The rest of this paper is organized as follows. Section II reviews the current state of research on IBD analytics and security technology. In Section III, a research framework for privacy-preserving and secure IBDA is proposed, which includes a platform architecture and three key technologies. Section IV outlines the IBDA platform architecture. Section V

discusses the key technologies that enable secure and efficient IBDA. The paper concludes with a summary in Section VI.

II. A Survey

This paper aims to provide guidance and references for research on privacy-preserving and secure IBDA. According to the previous analysis, IBD presents new characteristics such as multi-source, heterogeneity, and multimodality. Privacy leakage is a critical issue in the process of IBD sharing and trading among different parties in the open Industrial Internet environment. Moreover, the lack of mutual trust among data providers and consumers poses a significant challenge for data circulation and value creation. The current state of research on existing *big data fusion and analytics* methods and *data privacy and security protection* approaches, as well as their applicability in IBD scenarios, requires further investigation. In addition, *blockchain* technology, which can address the mutual trust issues, should also be taken into account.

A. Big Data Fusion and Analytics

1) Data Fusion: In industrial data analytics, integrating heterogeneous data from multiple sensors can significantly enhance the stability and reliability of the analytics model. Multi-modal fusion methods can be classified into three types based on the stage of data fusion: early fusion, late fusion, and hybrid fusion.

Early fusion integrates multi-source or multi-modal data into a single feature vector, which is then used as input for machine learning (ML) algorithms. Representative methods include TFN [37] and LMF [38] which are based on matrix operation. However, these methods have some disadvantages, such as complex computation and difficulty in extracting high correlations between modalities at the feature and data layers.

Late fusion, also known as decision-level fusion, involves training ML models on different modal data and then fusing the output results of multiple models. Late fusion is widely used in industry due to its flexibility and independence from feature extraction. Currently, late fusion primarily employs rule fusion methods such as maximum fusion, average value fusion, Bayesian rule fusion, and ensemble learning.

Hybrid fusion combines the advantages of both early and late fusion, while also increasing the structural complexity and training difficulty of the model. It is widely used in fields such as multimedia, visual question answering, and gesture recognition [39]. Zadeh *et al.* [40] used a memory fusion network and a special attention mechanism to simultaneously capture temporal and inter-modal interactions to obtain better multi-view fusion features. Mou *et al.* [41] used an attention-based convolutional neural network (CNN) and a long short-term memory (LSTM) network to integrate various in-car data for driver pressure detection. Lv *et al.* [42] proposed a differentiated learning framework that fully utilizes the diversity between multiple modalities to achieve more effective cross-modal domain adaptation.

2) distributed computing and approximate computing: Big data is characterized by its large scale, dynamic changes and low reliability. In some practical big data applications, it can be difficult or unnecessary to obtain an optimal solution. As a result, pursuing approximate results that can be computed efficiently and meet the requirements is of great importance. Approximate computing and distributed computing are two important means to improve the efficiency of big data analytics.

Approximate computing can be applied at the data, query, and system levels. Data-level approximation involves transforming big data into smaller data while retainng its main features, with little or no impact on the accuracy of query results. This approach can be used for tasks such as shortest path calculation [43] and link prediction [44]. Query-level approximation involves transforming high-complexity queries into low-complexity queries with little or no impact on the accuracy of query results. This approach can be used for tasks such as graph pattern matching [45], trajectory simplification [46], and dense subgraph computation [47]. System-level approximation is implemented at both the hardware and software layers [48, 49], including programming languages, compilers, memory, and processors.

Distributed computing focuses on the division and aggregation of data and models in a distributed environment, with the goal of improving analytics efficiency through parallel computing across multiple working nodes [50]. MapReduce is a common analysis framework [51] that breaks down analysis tasks into mapping and reduction processes. Big data analytics based on MapReduce includes Spark MLlib [52] and Vowpal Wabbit [53]. Frameworks based on parameter servers logically separate working nodes from model storage nodes to support more flexible collaborative working modes. Influential parameter server systems include CMU's Petuum [54], and Google's DistBelief [55]. Data stream-based distributed computing frameworks describe computing as a directed acyclic graph, with typical examples including TensorFlow and Pytorch. These computing frameworks have been widely applied in IBD scenarios, such as production process monitoring [56] and traffic flow forecasting [57] based on MapReduce and Hadoop.

3) Big Data Intelligent Analytics: Big data analytics typically involves four main tasks: regression, classification, clustering, and association rule mining (ARM). Regression is used to predict continuous values and is commonly used for tasks such as fault detection and quality assessment in industrial scenarios. Common methods include Support Vector Regression (SVR) [58], Random Forest (RF) [59], and Neural Network (NN) [60]. Classification involves predicting discrete values or mapping data items to different categories. Classification algorithms such as NN, Support Vector Machine (SVM), Decision Tree (DT), K-nearest neighbors (KNN), and Naive Bayes (NB) are widely used in industry [61]. Clustering is used to determine categories for unlabled data and is widely applied in scenarios such as pattern recognition [62], yield analysis [63], quantitative evaluation [64], and equipment status diagnosis [65]. Common clustering algorithms in industry include K-means [62], hierarchical clustering [64], DBSCAN

[66], and Self Organizing Maps (SOM) [62]. ARM is used to discover relationships between items in large-scale datasets, with common algorithms in industry including Apriori, FP Growth, and Eclat [67].

Data fusion, distributed approximate computing and intelligent analytics have seen rapid development in recent years and have been applied in IBDA. However, existing research primarily focuses on data analytics in specific industrial scenarios, with a lack of research on the fusion and analytics of multi-source, heterogeneous, and multi-modal IBD. When traditional data analytics models are directly applied to industrial scenarios, the domain adaptation, interpretability, and generalization ability of the model cannot be guaranteed.

B. Data Privacy and Security Protection

1) privacy protection: Existing privacy protection technologies include differential privacy (DP), homomorphic encryption (HE), secure multi-party computation (SMPC), and federated learning (FL).

DP is a method for protecting users' privacy information in published data by adding noise to the data. It has strict definitions and constraints, and aims to completely eliminate the possibility of privacy leakage from the data source [68]. In terms of big data privacy protection, Du *et al*. [69] explored the problem of differential privacy protection for training models in wireless big data scenarios, while Zhou *et al*. [70] focused on privacy protection methods for online social multimedia big data.

HE allows calculations to be performed on encrypted data, with the decrypted results matching those performed on the plaintext. In terms of big data privacy protection, Li *et al*. [71] used HE to solve the security data processing of industrial internet of things applications, protecting the privacy between data owners, untrusted cloud servers, and data users. Lu [72] focused on secure data queries based on HE to achieve privacy protection in a fog computing environment.

SMPC can ensure that participants' privacy information is not exposed in distributed scenarios. It is achieved through the use of various modern cryptography technologies, such as garbled circuit, oblivious transfer, secret sharing, HE, zero-knowledge proofs. Current mainstream computing frameworks include ABY, SPDZ, PICCO, and Obliv-C [73–75].

FL allows large-scale participants to perform distributed machine learning or deep learning while keeping data local, protecting privacy through encrypted parameter exchange [76]. Although FL does not directly expose participants' training data, research shows that it is still vulnerable to various privacy theft attacks. They include model extraction attacks that use the model service interface to predict specific parameters and the architecture of the privacy model; model inversion attacks that reconstruct model input through class tags returned by the model and confidence coefficients; membership inference attacks oriented to specific information in the training data set; and adversarial training attacks targeted at the model training process [77]. To address these attacks, feasible defense schemes such as SMPC-based secure aggregation [78], DP-based privacy training [79], secure training methods based on trusted execution environment (TEE) [80], and blockchain-based secure FL architecture [81]. For example, Bonawitz *et al*. [78] proposed a secure aggregation model that combines techniques like secret sharing to prevent the server from decrypting individual client gradients, thus hiding information from malicious servers. Approximate DP techniques such as Bayesian DP and centralized DP can further balance privacy and model availability [82, 83]. To hide local models and prevent theft of local model training results by other clients or servers, users can train their local models in a TEE with cryptographic protection. Kim *et al*. [84] proposed the BlockFL architecture, which uses a blockchain system to exchange and verify local learning model updates, with both local and global model updates added to the distributed ledger as blocks.

2) Data Security: Data encryption and access control are two crucial research areas for ensuring the security of industrial big data.

Data encryption: Big data encryption methods mainly include fully homomorphic encryption (FHE), convergent encryption (CE) and searchable encryption (SE). FHE allows for additions and multiplications on ciphertext any number of times, effectively reducing the risk of plaintext being stolen, intercepted, altered, or forged during transmission [85]. CE uses the hash value of data to encrypt it, reducing storage overhead and allowing for deduplication in the encrypted state [86]. SE enables keyword search on ciphertext without revealing any useful information about the plaintext [87].

Access control: In the IBD environment, cross-domain access and collaborative work are two major characteristics. For distributed big data environments, access control methods based on the attribute-based encryption (ABE) algorithm have been proposed in [88], which can provide fine-grained access control to some extent, but cannot dynamically evaluate the behavior of access subjects and grant minimum permissions. Zero-trust theory [89–91] offers a solution to these problems. Tao *et al*. [89] proposed a zero-trust-based secure analytics method for big data that can identify and intercept risky data access. Zaheer *et al*. [90] proposed a micro-service security system guided by zero-trust, emphasizing the use of zero-trust to improve security on the critical path of data flow. Chen *et al*. [91] designed a zero-trust architecture for smart healthcare platforms to achieve fine-grained access control.

Current privacy protection methods do not effectively balance the privacy requirements of IBD with model efficiency. A more efficient and secure privacy protection mechanism is needed to address privacy attacks and model poisoning attacks on IBD. Additionally, existing access control mechanisms cannot meet the needs of fine-grained access control and dynamic continuous trust evaluation in the complex access scenarios of the Industrial Internet.

C. Blockchain for IBDA

FL and DS&T are the two key applications of blockchain in IBDA scenarios. The research results in these two areas can be summarized as follows.

1) Blockchain for FL: The combination of blockchain and FL can enable decentralized model aggregation. Multiple

smart contracts can be used to securely verify and exchange local model updates. Qu *et al*. [92] and Unal *et al*. [93] used blockchain technology to address data poisoning attacks in FL systems. Some researchers have also used blockchain to enhance the privacy of FL. Jia *et al*. [94] developed a blockchain-based FL system to achieve multiple levels of data protection. Gao *et al*. [95] proposed a privacy-preserving asynchronous FL framework based on blockchain, which ensures trustworthiness by storing local models in the blockchain and generating the global model using a consensus algorithm.

The blockchain system has helped solve the problem of insufficient model stability and reliability in the FL process, while also improving training efficiency. Rehman *et al*. [96] proposed the concept of reputation-aware fine-grained FL based on blockchain, ensuring reliable collaborative training of data models. Moudoud *et al*. [97] used a blockchain sharding mechanism to improve blockchain efficiency and enable parallel model training, and designed a reputation mechanism based on multi-weight logic to incentivize data model updates. Kim *et al*. [98] proposed a local learning weighting method and a node selection method based on blockchain technology to improve the stability of FL.

In terms of consensus mechanism research, the latest few blockchain-based FL systems [99–102] still use basic PoW or PoS mechanisms. However, Li *et al*. [103] introduced an innovative committee consensus mechanism, where a group of honest nodes form a committee to verify local gradients and generate blocks. Additionally, Lu *et al*. [104] developed a training quality proof (PoQ) mechanism that combines data model training with consensus processes to better utilize the computing resources of nodes.

Regarding smart contract research, Rehman *et al*. [105], Mugunthan *et al*. [106], and Zhang *et al*. [107] developed smart contracts to maintain the reputation of FL participants and reward those who provide high-quality data. Martinez *et al*. [108] introduced a Class Sampled Verification Error Scheme, based on smart contracts, to verify and reward participants for uploading model gradients. Lee *et al*. [109] used smart contracts to manage the authentication of FL participating nodes and the distribution of global or local models.

2) Blockchain for DS&T: The circulation and application of industrial data involve issues of data ownership and transactions. Zhao *et al*. [110] introduced a blockchain-based data transaction protocol that considers both data availability and the anonymity of data providers. Dai *et al*. [111] suggested transforming traditional data trading platforms' data hosting/exchange services into data processing services, where only analytics results are accessible to data brokers and buyers. However, this solution is only applicable to Ethereum and SGX platforms. Regarding consensus mechanism research, Cui *et al*. [112] established a trust model based on vehicle interaction in the context of Internet of Vehicles (IoV) data sharing and designed an enhanced Delegated Proof of Stake (DPoS) consensus algorithm based on the trust model to balance security and efficiency.

In terms of smart contract research, Hu *et al*. [113] and Zheng *et al*. [114] developed a distributed data transaction scheme and transaction reward allocation rules based on smart contracts. Jiang *et al*. [115] introduced a new data transaction scheme for the industrial internet of things, consisting of smart contracts for packet transactions and analysis services. Kang *et al*. [116] used smart contacts to design a vehicle data storage and sharing scheme. For automated analysis, Muchhala *et al*. [117] incorporated the MapReduce paradigm into smart contracts to achieve more secure, effective, and transparent concurrent data calculation and analytics.

There have been some achievements in blockchain-based FL and DS&T research. However, existing studies have only designed blockchain systems for specific application scenarios. From the perspective of building an IBDA platform, it is necessary to consider the requirements of both application scenarios and design a scalable block structure. Additionally, while some researchers have designed a unique consensus mechanism for specific needs, most blockchain systems still use PoW or PoS, which can lead to security and performance issues in IBDA scenarios.

III. Challenges and the Research Framework

A. Challenges

To fully realize the value of IBD, it is necessary to promote the development of the IBDA platform, create new application modes, conduct extensive data sharing, and perform deep integration analytics. However, due to the new characteristics of IBD and the complexity of the Industrial Internet environment, secure and efficient IBDA faces the following challenges:

Challenge1: How the data intelligent analytics model fit into the new characteristics of IBD

IBD is derived from the different production processes of multiple industrial production enterprises across various industries. It has distinct characteristics such as being multi-source, multi-modal, large-scale, spatio-temporally correlated, and having strong scene differences. Existing IBDA studies primarily focus on single-scale industrial data analytics within specific scenarios of individual enterprises. They fall short of fully revealing the multi-scale correlation characteristics of multi-source, heterogeneous, and cross-domain IBD that are prevalent in the Industrial Internet. And they also fail to fully exploit the fusion effect of multi-source data. The model's usability is limited by its lack of domain adaptation, weak generalization ability in complex and dynamic industrial scenarios, and lack of interpretability for decision results.

Challenge2: How to analyze and utilize data while ensuring privacy and security

IBD has intricate ownership, circulation paths, and access relationships. Sharing and trading data may expose business secrets or personal privacy. The privacy protection demands of multi-source IBD vary, requiring IBDA platforms to identify sensitive attributes and provide hierarchical protection. FL, a new distributed machine learning paradigm, allows for multi-party joint modeling without exposing local data, greatly protecting data privacy. However, issues such as model privacy attacks, inefficient computing and communication, and dishonest participants in FL still pose challenges for data collaborative analytics. Additionally, the platform should

support fine-grained access control to handle the dynamic change of trust due to the complex ownership and access relationships of IBD.

Challenge3: How to share and trade data efficiently in the Industrial Internet environment without mutual trust

The IBD value is unclear, and relevant data trading standards are not well-established. This results in enterprises and users being unable, undaring, or unwilling to share or trade their data. Poor participation and difficulties in data provenance and auditing severely hinder DS&T. Blockchain technology offers a practical solution to these issues. However, to meet the demands for FL and DS&T, it is necessary to address the challenges of multi-source heterogeneous data storage and establish reliable mechanisms for managing the reputation of data subjects, as well as setting up rewards and penalties.

B. The Research Framework

In the construction of the Industrial Internet, a key issue is how to effectively analyze and utilize IBD while ensuring privacy and security. The analytics and utilization of IBD is a systematic process that requires the design of infrastructure that takes privacy and security protection into consideration, the organic integration of different security technologies, and coordination with big data analytics technologies. This paper proposes a research framework for privacy-preserving and secure IBDA, as shown in Fig. 1, which includes four research directions: the IBDA platform architecture, elastic fusion and analytics methods of IBD, privacy and security protection methods of IBD, and blockchain technologies supporting secure and efficient IBDA. The three objectives correspond to the three challenges presented in Section III-A. Obj1 is to develop novel methods for analyzing IBD, taking into account specific IBD characteristics and environments. Obj2 is to address the security and privacy problems in the process of big data analysis, focusing on FL and dynamic access control. The proposed FL framework aims to ensure privacy preservation in the distributed computing and cross-domain analysis processes involved in Obj1. Obj3 is to enhance the mutual trust in DS&T, improve the security of FL involved in Obj2 using blockchain technology, and uniformly store IBD features and fusion analysis results obtained in Obj1 using a scalable block structure. Together, these three objectives achieve the full exploitation of the combined value of IBD through DS&T under the premise of ensuring security and privacy.

1) IBDA Platform Architecture: Firstly, a thorough analytics of the Industrial Internet big data assets and their attributes is required. And the requirements of data collection, storage, analytics, sharing, and trading processes need to be considered to build the function model of the IBDA platform. Then, the security architecture and system architecture of the IBDA platform can be built by analyzing the privacy and security protection demands throughout the entire IBD lifecycle and integrating a variety of security and privacy technologies.

2) IBD Fusion and Analytics Methods: In terms of data representation, the joint representation method of multi-source data based on the multi-scale heterogeneous graph model can be studied to provide basic support for data fusion and intelligent analytics, aiming at the spatio-temporal correlation features of IBD. In terms of fusion analytics, the multi-modal fusion analytics model based on reliability measurement, differentiated learning, and the self-inductive learning framework needs to be studied to improve the domain adaptation and interpretability of IBDA. In terms of computing, the distributed elastic computing method for IBDA needs to be researched to support the elastic expansion of data, tasks, and models.

3) IBD Privacy and Security Protection Methods: Firstly, in light of the various sensitivity levels of IBD, the identification methods of sensitive attributes need to be studied. Then, to address the issues of model privacy attacks and dishonest clients, a secure FL framework enhanced by blockchain and TEE must be researched to support the training of the IBD intelligent analytics model with distributed and privacy protection. Finally, to address the issues of multiple access subjects and complex interaction scenarios in the Industrial Internet, the dynamic trust evaluation technology of access subjects can be studied using zero-trust theory. A hybrid role-attribute access control model based on trust evaluation needs to be proposed.

4) Blockchain Supporting Secure and Efficient IBDA: Firstly, a demand analysis for blockchain technology in IBD applications is required, and a scalable block structure supporting data storage and retrieval in key application scenarios needs to be studied. Then, to meet the specific characteristics and requirements of secure FL and DS&T, the reputation proof consensus algorithm and the hybrid consensus algorithm with high throughput and low cost need to be studied, respectively. Finally, multiple smart contracts needs to be studied to meet the security and efficiency requirements of various applications in the IBDA platform.

IV. IBDA Platform Architecture

In this section, the function model, security architecture and system architecture of the IBDA platform are proposed.

A. Function Model

IBD encompasses data from the entire industrial production life cycle, including enterprise informatization data, industrial internet of things data, and external cross-domain data. Common IBD assets are listed in Table I. IBD is dispersed across various stages of industrial production, originating from different systems and even different enterprises. Its potential value can be explored through sharing and trading. With extensive sharing and trading, IBDA can help enterprises make better decisions, gain deeper insights into customer behavior and preferences, and maintain competitive advantages in a fierce market environment. Therefore, we propose a function model for the IBDA platform, as shown in Fig. 2. This model is divided into application layer, data processing layer, cloud infrastructure layer, and data access layer, covering multiple industries such as energy, transportation, healthcare, chemical, and manufacturing.

1) The data access layer provides functions such as data access, data preprocessing, data classification, and data provenance. Firstly, the data access function supports the IBDA

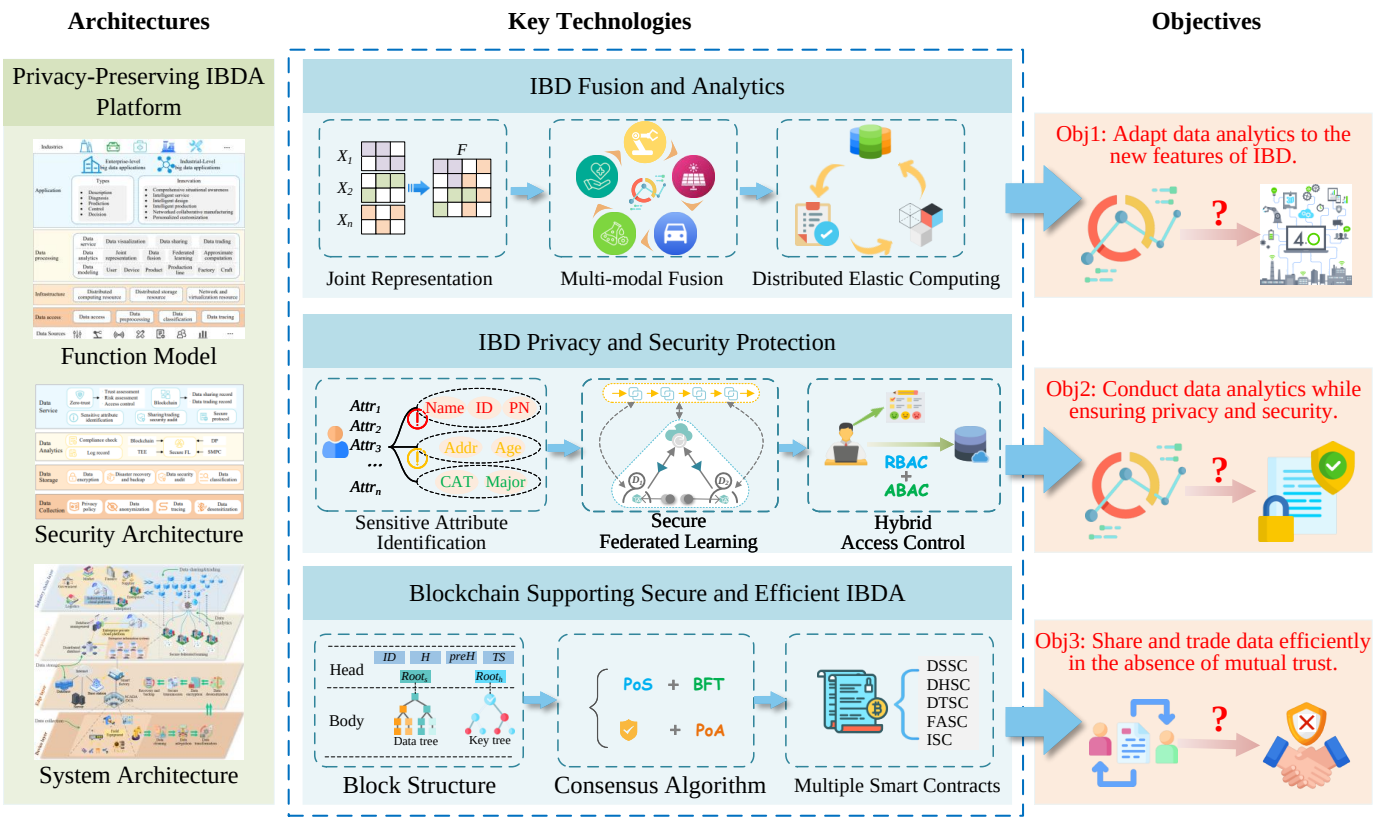

Fig. 1. Research framework for privacy-preserving and secure IBDA.

TABLE I
COMMON IBD ASSETS

Category	Type	Typical source	Data structure	Real-time
enterprise informatization data	design data	product model drawing document	semi-structured/non-structured	no
	value chain management data	SCM CRM	structured/semi-structured	no
	resource management data	ERP/OA MES PLM warehouse management system energy management system	structured	no
IIoT data	ICS data	DCS PLC	structured	yes
	production monitoring data	SCADA	structured	yes/no
	sensor data	external sensor barcode RFID	structured	yes
	other external device data	camera microphone	non-structured	yes
external cross-domain data	external data	related industry data regulation market data financial data logistics data environmental data	non-structured	no

platform in accessing data from the industrial internet of things, enterprise informatization systems, and external cross-domain systems. Secondly, the data preprocessing function offers capabilities such as data cleaning, conversion, and integration to preprocess the original IBD, eliminating incorrect and duplicate data to improve its effectiveness. Thirdly, the data classification function provides differentiated protection and management for data of various types, security levels,

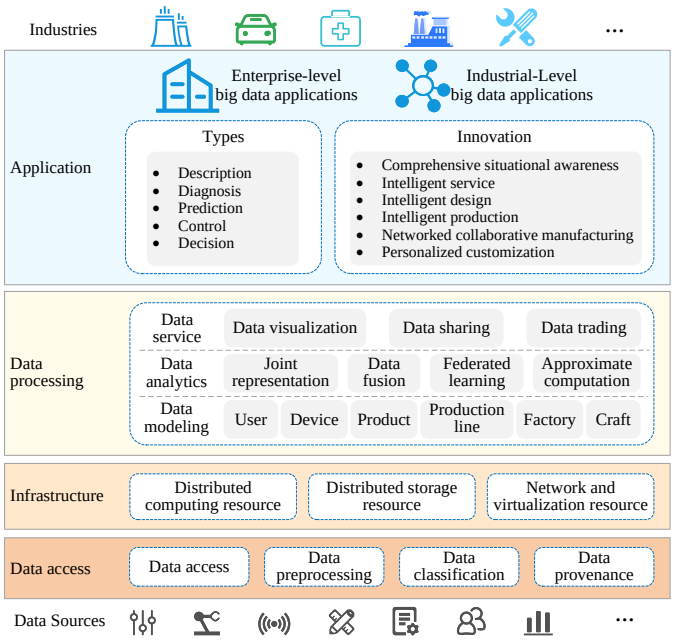

Fig. 2. Function model for the IBDA platform.

and privacy protection levels. Fourthly, the function of data provenance supports data accountability and authenticity verification.

2) The infrastructure layer offers infrastructure for data analytics, sharing, and trading. This includes resources for distributed computing, distributed storage, networks, and virtualization.

3) The data processing layer provides functions such as data modeling, data analytics, and data service. Digital models such as users, equipment, products, production lines, factories, and processes are established based on the strong mechanism of industrial data. The data processing layer offers approaches such as multi-source data joint representation, multi-modal data fusion, FL model training, and approximate computing. Big data and artificial intelligence technologies are used to deeply mine and analyze the potential value of IBD. This layer also integrates data services such as data visualization, data sharing, and data trading.

4) The application layer provides traditional types of big data applications such as descriptive analysis, diagnostic analysis, predictive analysis, control analysis, and decision analysis. In addition, it also includes innovative big data applications that support enterprise- and industrial-level comprehensive situational awareness, intelligent service, intelligent design, intelligent production, networked collaborative manufacturing, and personalized customization.

B. Security Architecture

The analytics and utilization of IBD includes five processes: collection, storage, analytics, sharing, and trading. The security and privacy of data must be ensured in each process to realize the data security and privacy protection of the entire IBDA platform. Fig. 3 summarizes the security and privacy protection requirements for the entire process of IBD analytics and utilization.

We propose a security architecture for the IBDA platform, as shown in Fig. 4. The architecture consists of four layers: *data collection*, *data storage*, *data analytics*, and *data service*. Each layer addresses the security and privacy protection requirements of the corresponding processes in Fig. 3.

1) The data collection layer ensures that the data owner is informed and consented to the data collection behavior, and performs autonomous collection based on the minimization principle declared in the privacy policies provided by the requesting party. Data anonymity and desensitization techniques are applied to protect basic privacy information. Data provenance techniques guarantee the authenticity and traceability of data by recording data collection logs, including data sources, collection times, and data summaries.

2) The data storage layer employs data encryption, disaster recovery and backup, and data security audit measures to ensure the confidentiality, integrity, reliability, and availability of data. Data should be classified and stored according to factors such as data security level, privacy level, importance, and frequency of use.

3) The data analytics layer first needs to verify the compliance of the analytics task, including verifying whether the data source, analytics purpose, and analytics logic are legitimate, whether privacy data is involved, whether the analytics results are consistent with the statement, and whether the use of the analytics results is compliant. Then, the data analytics layer provides a secure FL framework, combining blockchain, TEE, DP, and SMPC technologies to realize privacy-preserving data analytics. The data analytics process is recorded in the form of a log.

4) The data service layer involves security measures for data sharing and trading. Sensitive attribute identification technique is applied to analyze the attribute sets that may compromise privacy in order to support the settings of classified access policies. The security audit of data sharing/trading verifies the legality and compliance of data users' requests, and enforces the minimization principle during data sharing. Blockchain is utilized to track and record data flow information, including logs of data sharing and trading, thereby supporting traceability and trade supervision. Access control follows a zero-trust concept, dynamically adjusting the authority of data users through trust evaluation of the subjects and risk assessment of their behavior in sharing or trading data. Security protocols are employed to guarantee confidentiality and integrity during communication.

C. System Architecture

To illustrate the deployment and implementation of the IBDA platform in the industrial production process, as well as the application of key security and privacy technologies in the platform, we propose a system architecture for privacy-preserving and secure IBDA platform, as shown in Fig. 5. This architecture includes four layers: device layer, edge layer, enterprise layer, and industry chain layer. It covers the collection, storage, analytics, sharing, and trading of IBD.

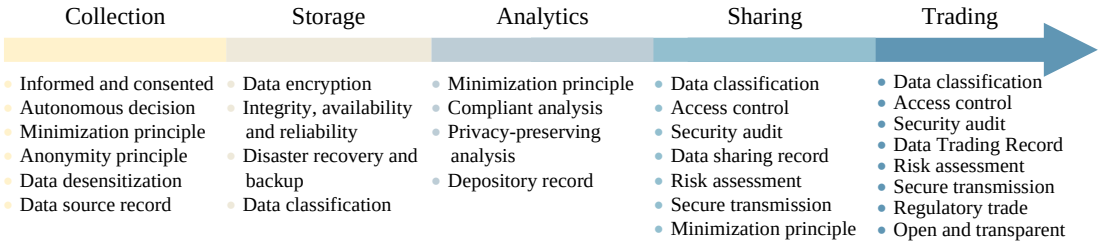

Fig. 3. Security and privacy protection requirements of the whole process of IBD analytics and utilization.

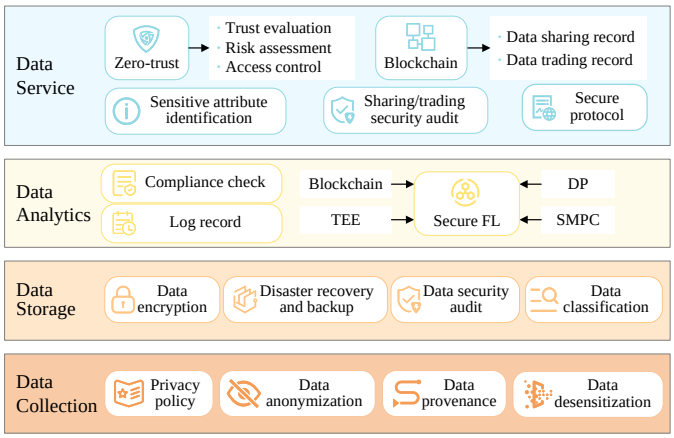

Fig. 4. Security architecture for the IBDA platform.

1) The device layer serves as a significant data source for the IBDA platform and must support the collection of data from a wide range of industrial control field devices with complex protocols. This includes the collection of massive, real-time/batch, structured/unstructured, and sequential/non-sequential data. The device layer also provides basic data pre-processing operations such as data cleaning, data integration, and data conversion.

2) The edge layer deploys monitoring and control systems that directly interact with industrial control field devices, as well as communication, computing, and storage infrastructure. The edge layer provides edge-side data storage capability for industrial production fields and supports high-concurrency, high-throughput, and high-performance data storage.

3) The enterprise layer deploys various information systems, distributed databases, and enterprise-level private cloud platforms of industrial enterprises. The enterprise information system is responsible for internal management functions such as supply chain management, product life cycle management, and customer relationship management. Distributed databases support the local storage of IBD within enterprises and enable secure and controllable autonomous management. Enterprise private cloud platforms perform intelligent data analytics and data sharing within enterprises to drive intelligent development. At the same time, these platforms also participate in industry-level DS&T and use enterprise local data to participate in the construction of industry-level intelligent analytics models under the framework of secure FL.

4) The industry chain layer deploys industry-level public cloud platform, which optimizes industrial resource allocation and creates a creative industrial ecology through data analytics, sharing, and trading between enterprises. Data analytics is carried out based on the secure FL framework, allowing all enterprises to cooperate to complete the ML model training task of the data analytics sponsor. The parameters and model exchange during model training are realized based on the industry-level blockchain system. DS&T within the industry chain are also carried out based on the industry-level blockchain system, with smart contracts utilized to complete DS&T recording.

V. Key Technologies

A. IBD Fusion and Analytics Methods

The multi-source heterogeneous IBD representation method, the multi-modal fusion analytics method, and the distributed elastic computing method are presented in this section.

1) Multi-source Heterogeneous IBD Representation Method: The representation of multi-source heterogeneous IBD includes single modal IBD representation and multi-modal IBD joint representation.

a) Single Modal IBD Representation Method Based on Multi-scale Deep Model: Industrial data comes in many forms, but most industrial production data has either temporal or spatial attributes. By observing this data at multiple temporal and spatial scales, we can uncover data patterns at different levels to improve the accuracy of downstream data analytics. In our previous work, we proposed extracting multi-scale temporal features through wavelet analysis and fusing multi-scale features through attention mechanisms for temporal data[118]. For spatial data, we proposed using Convolutional Neural Networks (CNN) to obtain multi-scale spatial features[119]. And in [120], we proposed a nested attention mechanism for multi-scale image information fusion, which can be further extended to general spatio-temporal data.

For more general industrial data $D = \{t_i, (x_i, y_i), a_i\}$, with both temporal and spatial attributes, we suggest using a multi-scale deep model for joint modeling of spatio-temporal information. Among them, t_i represents time, (x_i, y_i) is a two-dimensional spatial coordinate, a_i represents the specific measurement value at time t_i at position (x_i, y_i). Spatio-temporal data is expressed as a snapshot sequence after continuous time is discretized according to a specific temporal scale. The data value at a particular spatial location on the snapshot is taken as the average of observations at adjacent times at that location. The analogy can be made that each snapshot corresponds to a single frame of an image, while multiple snapshots along

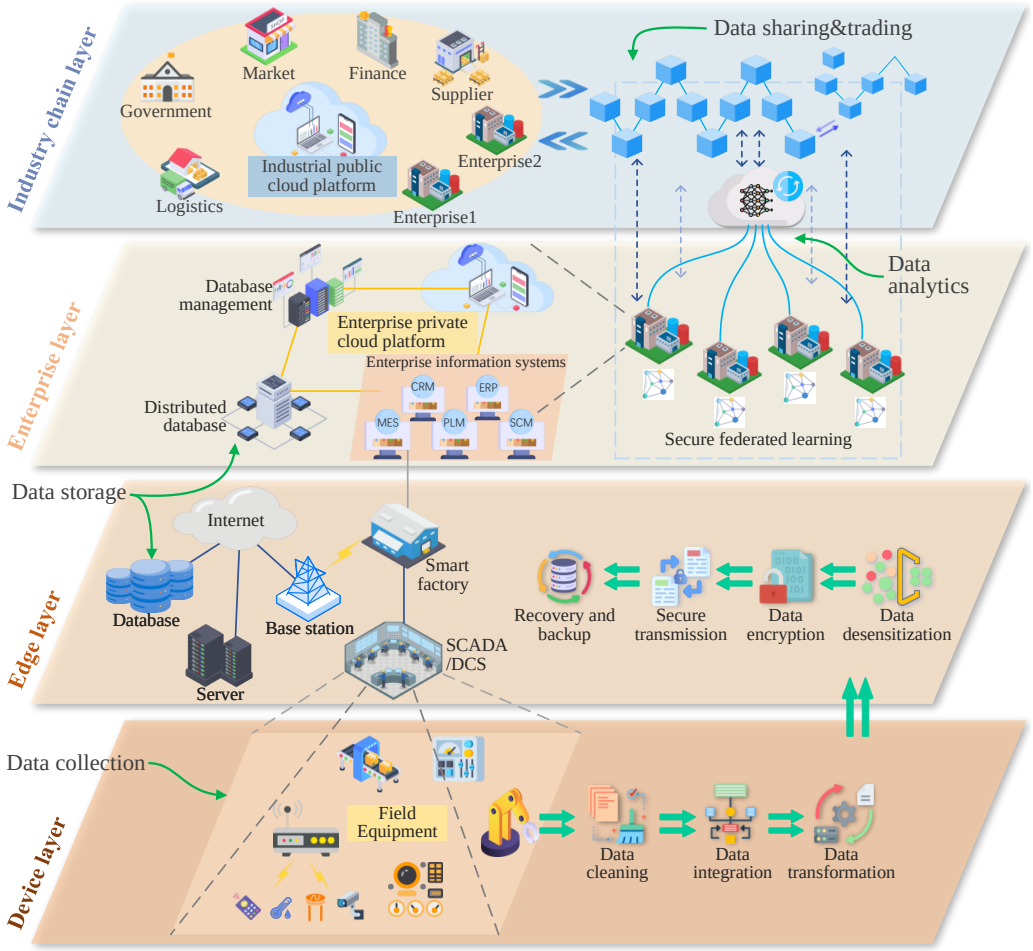

Fig. 5. System architecture for privacy-preserving and secure IBDA.

the timeline are equivalent to consecutive frames in a video. Therefore, multiple layers of CNN can be used to obtain depth features at multiple spatial scales for each snapshot, with the resulting feature vector input into LSTM, forming a typical CNN+LSTM architecture. To increase the model's sensitivity to spatial position, precise position information, such as normalized longitude and latitude, can be embedded in each grid involved in the convolution, and a position-sensitive pooling algorithm can be used.

To extract depth features from multiple temporal scales, the time interval between snapshots can be adjusted to obtain spatio-temporal sequences at different temporal scales. Depth features at different spatial scales can be extracted using CNN. After obtaining spatio-temporal features at different scales of a single modal data, the multi-scale features are fused through a nested attention mechanism. The complete steps of this method are as follows:

Step 1: Discretize continuous spatio-temporal data into time snapshot sequences $\{D_t\}$ at a certain temporal scale t;

Step 2: Obtain spatio-temporal data sequences $\{D_{t,s}\}$ by sampling on each time snapshot at a certain spatial scale s;

Step 3: Extract the depth feature $f_{t,s,j}$ of $\{D_{t,s}\}$ using CNN-LSTM model. $f_{t,s,j}$ represents the j-th feature at the temporal scale t and spatial scale s;

Step 4: Extract the depth feature set F at multiple spatio-temporal scales by adjusting t and s:

$$F = \{f_{t,s,j} | t \in T, s \in \Omega, j \in Z^+\}, \tag{1}$$

where T is the set of temporal scales, Ω is the set of spatial scales, and Z^+ is the set of positive integers;

Step 5: Use nested attention mechanism for spatial feature fusion at the same temporal scale:

$$f^{'}_{t,s} = Attention(f^{'}_{t,s-1}, f_{t,s})(1 < s \leq |\Omega|); \tag{2}$$

Step 6: Obtain spatial fusion feature at each temporal scale:

$$F^{'}_{t} = f^{'}_{t,|\Omega|}; \tag{3}$$

Step 7: Use nested attention mechanism for multi-scale temporal feature fusion along the timeline:

$$F^{''}_{t} = Attention(F^{''}_{t-1}, F^{'}_{t})(1 < t \leq |T|); \tag{4}$$

Step 8: Finally, the spatio-temporal fusion feature is obtained:

$$F^{'''} = F^{''}_{|T|}. \tag{5}$$

b) Joint Representation Method for Multi-modal IBD Based on Multi-scale Heterogeneous Graph Model: On the basis of multi-scale depth features of single-modal industrial data, it is necessary to further reveal correlation features between multi-modal data. In our previous study, we proposed a heterogeneous graph embedding algorithm based on heterogeneous random walk [121] and a multi-scale homogeneous graph representation method based on the ant colony algorithm [122]. Heterogeneous graphs effectively preserve the association information between multiple entities, allowing for effective mining of deep relationships within the heterogeneous graph structure through heterogeneous random walk. This approach is well-suited for association feature mining in multi-modal IBD. Additionally, the graph embedding method based on ant colony accurately captures the hierarchical clustering structure of the graph and generates proper multi-level embedding vectors for nodes. By integrating and optimizing these two approaches and extending them to the scenario of heterogeneous graphs in IBD, we propose a joint representation method for multi-modal IBD based on a multi-scale heterogeneous graph model.

Fig. 6 is a schematic diagram of this method, which includes the following four steps:

Step 1: Feature extraction: The multi-scale fusion features of the original data are obtained through the method described above. Based on the extracted features, we define four types of association between data sources:

- Spatio-temporal relation. The data source is within a similar range in terms of both space and time.
- Logical relation. There is a relationship of logical dependency between data sources.
- Entity relation. Specific entities establish indirect relationships between data sources.
- Statistical relation. There is a statistical correlation between data sources.

Step 2: Heterogeneous graph construction: The heterogeneous graph is constructed among different data sources, with vertices representing specific data sources and edges representing the above associations between data sources.

Step 3: Heterogeneous graph embedding: For heterogeneous graphs, we suggest using the multi-scale ant colony algorithm to obtain a multi-scale graph pyramid. First, the ants walk randomly on the heterogeneous map and release pheromones. When an ant passes through the same node, it means that a loop in the graph is detected. Each time the ant detects a loop, it stops walking and releases pheromones on the edges of the loop as follows:

$$\Delta\rho_{ij} = \frac{1}{length(loop)}, \tag{6}$$

where ρ_{ij} represents the amount of pheromone released, and $length(loop)$ represents the length of the loop. Every time an ant walks, the probability of walking from any vertex u_i to u_j is defined as:

$$P(u_i \to u_j) = \frac{W_{ij}\rho_{ij}^{\alpha}}{\sum_k W_{ik}\rho_{ik}^{\alpha}}, \tag{7}$$

where W_{ij} represents the weight of the connecting edge, ρ_{ij} is the pheromone, and α is a hyper-parameter. The pheromone concentration on the edges of the heterogeneous graph reflects the closeness of the relationship between the vertices. After performing random walks with multiple ants, the edges with higher pheromone accumulation indicate that the vertices they connect form shorter loops, implying a stronger association. We can use this pheromone information to identify and merge related vertices, resulting in a simplified, heterogeneous graph. By repeating this process, we can construct a graph pyramid consisting of multiple heterogeneous graphs at different scales.

Step 4: Heterogeneous graph collapse: The graph embedding optimization method is applied to different subgraphs at different scales to obtain the embedding representations of each vertex. Then, the joint representation of the graph structure of each data source is obtained by concatenating the embedding representations of different subgraphs and performing PCA dimension reduction. The joint representation preserves the inherent data features of each source and integrates the association features of multi-source data. This provides a foundation for deep association analysis and joint deep reasoning of multi-source industrial data.

2) Domain Adaptive and Interpretable Multi-modal Fusion Intelligent Analytics Method: Domain adaptation is a common problem in IBD intelligent analytics models due to different application scenarios in specific industrial practices. When a model trained on the training set is transferred to a target industrial scene for actual deployment and use, the actual data feature distribution in the scene often differs significantly from the training set due to factors such as environment, equipment, and human behavior, resulting in decreased model performance. Artificial intelligence applications in industrial scenarios are often closely related to industrial production decision-making. So, studying model interpretability can make their behavior more transparent and trustworthy, helping decision-makers understand the model's results and influencing factors and avoiding potential decision-making risks. Fig. 7 shows a suggested domain adaptive and interpretable multi-modal fusion intelligent analytics method. Based on reliability measurement, dynamic weighted fusion of multi-modal IBD is performed, followed by incremental optimization of the model using an unsupervised incremental optimization method based on multi-modal differentiated learning with unlabeled data to enhance domain adaptation. Finally, within a prototype-based self-inductive learning framework, the logical rules of prototype combination are summarized and induced to generate decisions with strong explanatory power.

a) Multi-modal Data Fusion Method Based on Reliability Measurement: In our previous work [42], we proposed a reliability measurement method based on a single prototype to address the domain adaptation problem, and enhanced the domain adaptation capability of the fusion visual classification model on multi-modal datasets. We extend this method to multi-modal industrial data scenarios, and propose to measure the model's generalization ability for different modal data in the target scenarios based on prototypes, as shown in Fig. 7.

A prototype training task is introduced during pre-training on the training set, and the feature vector is mapped to the

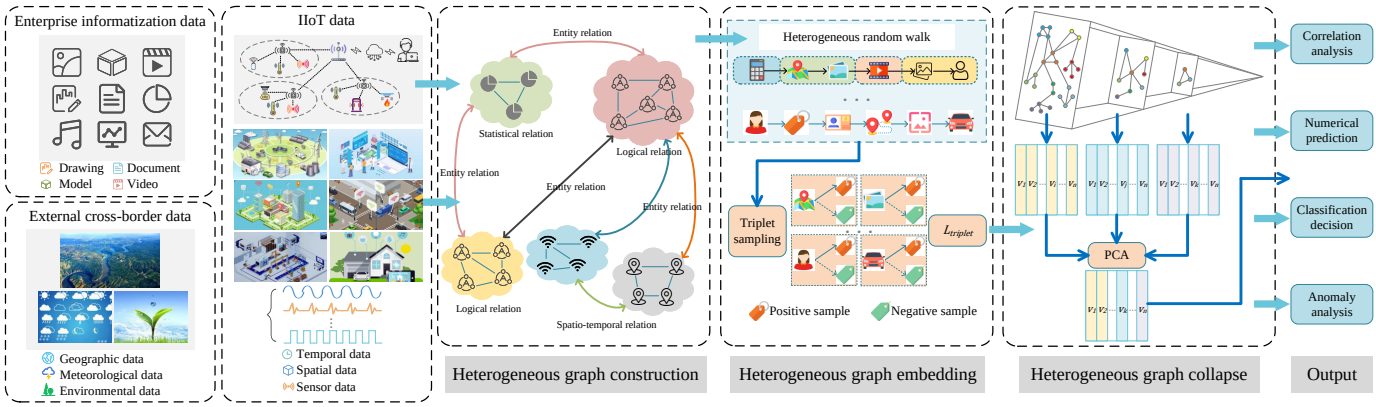

Fig. 6. Schematic diagram of the multi-modal industrial data joint representation method based on multi-scale heterogeneous graph model.

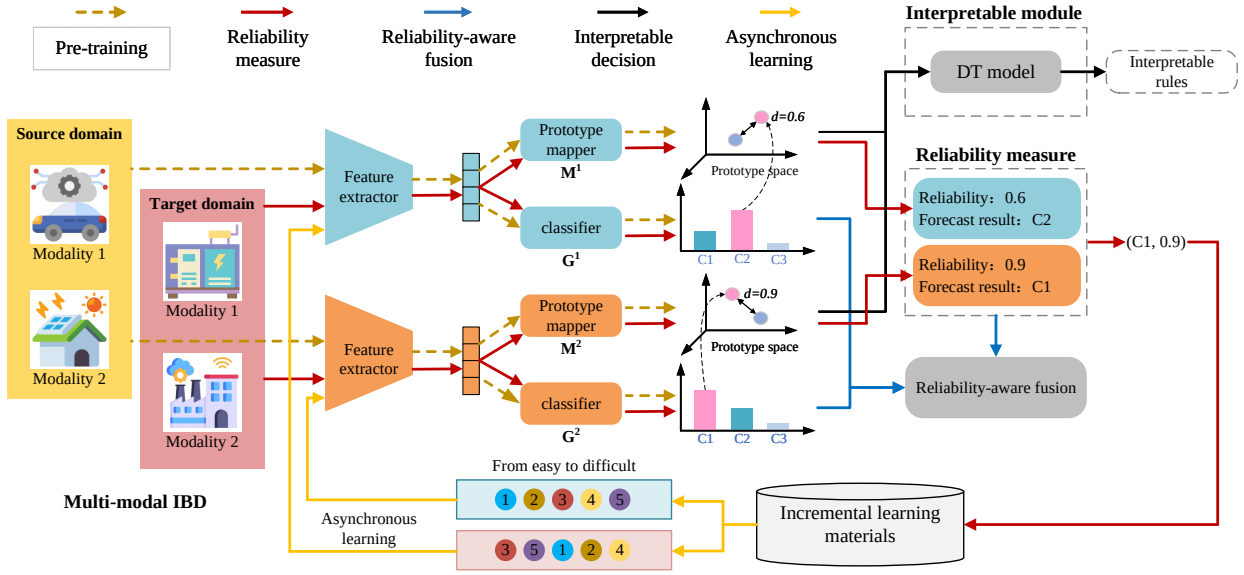

Fig. 7. Schematic diagram of a domain adaptive and interpretable multi-modal fusion intelligent analytics method.

low-dimensional prototype space using a prototype mapper. Random prototype vectors are assigned to each category in the prototype space, with the prototype vector of each category brought close to the sample vector of the same class by minimizing prototype contrast learning loss L_c.

$$L_c = -\sum_{q} log(\frac{exp(q \cdot W_{m,i}/\tau)}{\sum_{k}^{C} exp(q \cdot W_{k,j}/\tau)}), \tag{8}$$

where q represents the vector of any sample. m represents the class to which q belongs. $W_{k,j}$ represents the j-th prototype vector of the k-th class, randomly selected from the prototype vectors of the k-th class. $W_{m,i}$ represents the vector randomly selected from the prototype vectors of the m class, which is a positive sample. τ stands for temperature coefficient and is used to adjust the contrastive learning intensity of positive and negative samples. By combining L_c and the downstream learning task loss L_p, we obtain the following loss function:

$$L = L_p + \lambda L_c, \tag{9}$$

where λ is the hyper-parameter that balances the downstream task and the prototype contrastive learning task. By minimizing the joint learning loss L on the training set, we can obtain the prototype vector set $\{W_{k,j}\}$ that represents the typical sample features of each class k in the training set, while optimizing the feature extractor for downstream tasks.

When a trained classifier is transferred to a target scene, a prototype-based reliability measurement method is used to assess the model's domain adaptation. Specifically, the model's adaptation ability on a test sample is determined by comparing the difference between the feature vector and the prototype vector of the test sample. For a given sample x, if a classifier of a certain mode determines that the sample belongs to class m, then the reliability measurement function of this decision is defined as follows:

$$R(x, m) = \max_{j} a(\kappa(x, W_{m,j})), \tag{10}$$

where, $W_{m,j}$ represents the j-th prototype vector of the classifier in the class m. κ stands for kernel function and is used to measure the similarity between sample and prototype

vector. Some typical kernel functions available include Gaussian kernel, exponential kernel, inner-product kernel, etc.

Based on reliability measurement, the sub-model output decisions of each mode are weighted and fused according to their reliability measurements through late fusion, resulting in the final fusion decision:

$$\hat{Y} = \frac{\sum_{t=1}^{T} R^{(t)}(x, Y^{(t)}) \cdot Y^{(t)}}{\sum_{t=1}^{T} R^{(t)}(x, Y^{(t)})}, \tag{11}$$

where x is the input sample, $Y^{(t)}$ represents the classification decision of the classifier of the t-th modality for x, and $R^{(t)}(\cdot)$ represents the reliability measure of the classifier's decision.

b) Unsupervised Incremental Optimization Method Based on Differentiated Learning: As an effective method of incremental learning, curriculum learning can address the problem of single-modal domain adaptation. In our previous work [42], we extended curriculum learning to asynchronous curriculum learning for multimodal data classification. To further improve the adaptation of the multi-modal fusion analytics model to different industrial scenarios, we suggest using the multi-modal differentiated learning method for unsupervised incremental optimization of the model, as shown in Fig. 7. Firstly, sub-models with high reliability are selected for each sample in the target scene based on reliability measurement to obtain pseudo-labels, constructing a training data pool in the target domain. Then, an asynchronous curriculum learning mechanism is adopted, allowing different sub-models to select samples from the training data pool for training. The order of sample selection follows from easy to difficult, with difficulty measured according to reliability. Since different models have varying degrees of reliability on different samples, different sub-models will have different sample selection orders. This differentiated learning mode greatly improves the model's domain adaptation.

c) Prototype-based Self-induction Learning Framework: Existing research results indicate that prototypes are an effective way to improve interpretability of models [123], and that shallow classification models are more interpretable than deep models [124]. To achieve interpretability of the fusion analytics model, we can graft the shallow interpretation model into the reliability measurement process. By iteratively optimizing the fusion model and the interpretation model based on reliability measurement, the interpretable decision rules can be self-summarized at the same time as fusion reasoning. Our proposed prototype based self-inductive learning framework is shown in Fig. 7. We obtained prototype vectors of each mode in the aforementioned reliability measurement learning task. Each prototype vector represents a typical feature vector and corresponds to specific data space semantics. We can explain the prototype by combining experts' prior knowledge. These prototype vectors serve as basis vectors, spanning a multi-modal prototype space. The sample feature vector of each input is projected into the prototype space to obtain a low-dimensional semantic vector. Furthermore, a shallow model with strong interpretation is introduced, and the semantic vector is used as input to fit the final output of the fusion analytics model. Optional shallow models include KNN, logistic regression, and decision trees, with their output constituting interpretable decision rules.

3) Distributed Elastic Computing Method for IBD: The distributed elastic computing method for IBD comprises three aspects: an adaptive distributed approximate computing method for data elastic expansion, a hierarchical decoupling resource scheduling mechanism for task elastic expansion, and a pluggable intelligent component construction method for elastic deployment.

The distributed elastic computing method for IBD comprises three aspects: an adaptive distributed approximate computing method for data elastic expansion, a hierarchical decoupling resource scheduling mechanism for task elastic expansion, and a pluggable intelligent component construction method for elastic deployment.

a) Adaptive Distributed Approximate Computing Method: Compared to traditional approximate computing methods based on data synopses, the distributed approximate computing method for IBD must better adapt to dynamic streaming data due to IBD's characteristics of strong timing, multi-source heterogeneity, and massive dynamics. We propose improving the adaptive ability of distributed approximation algorithms in three ways:

- Data precision-aware adaptive adjustment. Each computing node uses non-uniform data synopses to conduct real-time statistics on incoming data. Moreover, the node adaptively adjusts the statistical accuracy of data synopses in different data intervals based on the degree of conflict (the number of data samples with the same approximate value) to improve the approximation accuracy of overall aggregation calculations.
- Task-aware adaptive adjustment. Create task-specific data synopses pools for different downstream approximate computing tasks, and perform adaptive elastic expansion as the tasks expand.
- Device capability-aware adaptive adjustment. Computing tasks are adaptively divided according to the computing power of nodes, with stronger computing nodes being assigned more tasks.

b) Hierarchical Decoupling Resource Scheduling Mechanism: IBD mining is a typical computation-intensive, delay-sensitive and communication-intensive application. In a traditional centralized scheduler, task scheduling and state information maintenance are coupled, with the global state of nodes and tasks being maintained during task scheduling, leading to a performance bottleneck.

We recommend using a hierarchical decoupling resource scheduling mechanism to decouple a centralized scheduler into a global state manager and a distributed scheduler. The global state manager maintains the status of nodes and tasks, while the distributed scheduler has two tiers: a global scheduler and a local scheduler. To avoid overloading the global scheduler, local tasks are scheduled locally first. When the local node is overloaded or cannot meet task requirements, the tasks are sent to the global scheduler for scheduling. The global scheduler schedules tasks based on the global state. When the global scheduler becomes a performance bottleneck, multiple copies can be quickly instantiated, share state information through

the global state manager. This allows for efficient and flexible expansion of a distributed scheduling system in the face of numerous IBD mining tasks.

c) Pluggable Intelligent Component Construction Method: As deep learning continues to develop, the paradigm of intelligent analytics methods is becoming increasingly similar. The core demand for industrial intelligence upgrading is to flexibly and effectively apply existing intelligent analytics technologies to diversified IBD mining scenarios, providing rapid secondary development and flexible deployment.

We recommend a pluggable intelligent component construction approach to facilitate the rapid integration and validation of intelligent analytics models for IBD mining scenarios. Firstly, according to different stages of intelligent analytics, common algorithms are modularized into reusable general components, including data preprocessing components, feature extraction components, downstream task components, and joint optimization components. Corresponding component interface specifications are formulated. Then, a component intermediate description language and corresponding compiler are designed to translate the component intermediate description language into an executable program. Finally, by implementing a visual AI component repository, users can flexibly combine different components in the repository through drag-and-drop functionality. The composition between components is translated into the component intermediate description language and ultimately compiled into an executable model for deployment in a distributed computing environment.

B. IBD Privacy and Security Protection Methods

In this section, we discuss the IBD sensitive attribute identification method, the IBD privacy-preserving computing method based on secure FL, and the hybrid access control model.

1) IBD Sensitive Attribute Identification Method: Data within the Industrial Internet exhibits varying levels of sensitivity due to differences in industry, source, and generation mode. As such, data storage and management require classification and grading. A feasible approach involves utilizing information entropy to quantify the sensitivity of data attributes and identifying sensitive attributes based on association rules [125]. Fig. 8 shows the overall process of this method.

Information entropy quantifies the degree of disorder of an attribute. A larger information entropy value indicates a higher level of disorder in the attribute's value. Maximum discrete entropy measures the maximum uncertainty associated with an attribute and represents the maximum effective information that an attacker can obtain after accessing attribute data. Thus, the sensitivity of an attribute can be determined by comparing the difference between its maximum discrete entropy and its information entropy. The sensitivity of the attribute A_i in IBD can be defined as:

$$S_i = \frac{H_{max}(A_i) - H(A_i)}{H_{max}(A_i)}, \tag{12}$$

$$X_t \in R^{M \times N \times C}, \tag{13}$$

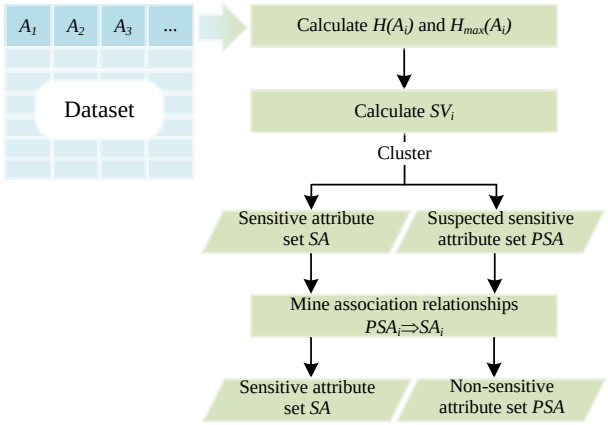

Fig. 8. The process of sensitive attribute identification.

where $H_{max}(A_i)$ is the maximum discrete entropy of attribute A_i, and $H(A_i)$ is the information entropy of attribute A_i. S_i represents attribute sensitivity, $S_i \in (0, 1)$. The smaller the value, the more sensitive the attribute is.

Upon quantifying the sensitivity of each attribute in a dataset, potential sensitive attributes can be identified by mining the association relationships between attributes. This approach serves to prevent attackers from inferring sensitive attributes from non-sensitive ones. According to the sensitivity calculation, attributes of data samples are preliminarily divided through cluster analysis to obtain sensitive attribute set SA and suspected sensitive attribute set PSA. Then, the Apriori algorithm is used to mine the association relationships between SA and PSA in the dataset, and all strong association rules are obtained. Further classification of the suspected sensitive attribute set can be determined by the number of strong association rules successfully established between SA and PSA. Finally, all attributes of the sample can be divided into sensitive attribute set SA and non-sensitive attribute set NSA.

2) Privacy-preserving Computing for IBD Based on Secure FL: FL enables distributed modeling without revealing local data, providing significant protection for local data privacy. However, FL remains vulnerable to various privacy theft attacks, including model inversion attacks, membership inference attacks, adversarial training attacks, and poisoning attacks. To enhance its security and privacy protection capabilities, it is feasible to introduce technologies such as DP, HE, TEE, SMPC, and blockchain into FL. The blockchain-based secure FL framework (BCFL) is a current hot research topic. Using blockchain systems can provide capabilities such as secure model parameter exchange and verification, client reputation management, training rewards and punishments, and decentralization.

A secure FL framework based on TEE and blockchain is shown in Fig. 9. This framework utilizes blockchain to publish and retrieve data, as well as to provide a secure exchange of model parameters and updates between clients and servers. The training process is recorded in a distributed ledger for auditing and tracing purposes. Blockchain establishes an infrastructure for all TEEs, offering reliable communication channels. This architecture ensures that the plaintext param-

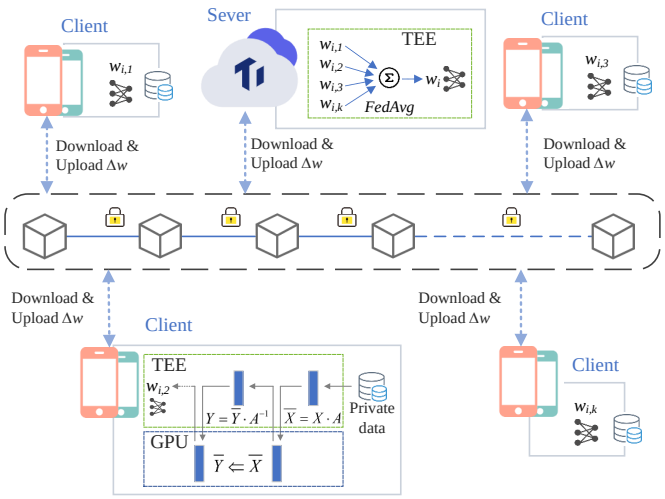

Fig. 9. Secure FL framework.

eters of the model are always kept within the TEE during local client training, providing defense against various model privacy attacks. Typically, all computing tasks associated with model training are performed within the TEE. If the client has access to a GPU or hardware accelerators, linear operations within the training task can be outsourced to the GPU, accelerating linear matrix operations. The secure FL scheme based on TEE and blockchain is as follows:

Step 1: The server-side TEE initializes the model to be trained and completes identity authentication and key negotiation with the TEEs of various federated clients through the blockchain. Identity and authentication information are stored on the chain.

Step 2: The server-side TEE encrypts the global model and uploads it to the blockchain. The client downloads the latest model parameters from the blockchain, and the TEE decrypts the model.

Step 3: The client-side TEE loads their respective local training datasets and performs forward and backward propagation computations within the TEE to obtain the model gradient. During the training period, computationally intensive matrix linear operations within the TEE can be outsourced to GPUs through secret sharing protocols, accelerating model gradient calculation.

Step 4: The client-side TEE executes FedAvg to update local model parameters. The updated model parameters are encrypted by the client-side TEE and stored in the blockchain.

Step 5: The server downloads model updates from the blockchain and decrypts them within the TEE. It then aggregates model parameters from all clients using FedAvg, updates the global model, and uploads it onto the blockchain.

Step 6: Repeat **Step 2-5** above until the global model converges.

Table II shows some application scenarios of the proposed secure FL framework in industry, as well as potential data, clients, and servers.

3) Dynamic Trust Evaluation Based on Zero-trust and Hybrid Access Control Model: In light of the complex data access requirements in IBD sharing and trading scenarios, it is suggested that dynamic trust evaluation of the subject, object, and environment in access scenarios be conducted based on the zero-trust concept. In conjunction with dynamic trust, research into a hybrid role-attribute access control model is recommended.

a) Dynamic Trust Evaluation Based on Zero-trust: In a zero-trust system, there is no traditional boundary trust mechanism, and no user, network or device is trusted by default. Each data access behavior of the subject must undergo dynamic trust evaluation, as shown in Fig. 10. Dynamic trust evaluation in the IBD environment accepts multi-source security elements from the subject, object and external environment. The subject can only access the target object if the dynamic trust evaluation results of the subject, object, and external environment meet the requirements. In dynamic trust calculation, required security elements are first collected and then cleaned, aggregated, labeled, and classified. The trust model can then be built by combining methods such as Bayesian probability analysis, outlier detection, peer group analysis, and fuzzy hierarchical analysis. Additionally, the trust value should be modified and updated based on historical trust value, recommended trust value of adjacent nodes, and trust feedback [126].

b) Hybrid Role-attribute Access Control Model: In the Industrial Internet, complex and diverse data sharing requirements pose great challenges to the efficiency and security of access control. Traditional access control models use static rules, blocklists, and allowlists for one-time evaluation, which can easily result in excessive authorization, data abuse, and privacy disclosure in a big data environment. Introducing a continuous trust evaluation mechanism into access control can enable dynamic risk perception. The traditional role-based authorization mechanism has the advantage of simplifying management. Therefore, we suggest a hybrid role-attribute access control model based on trust evaluation, as shown in Fig. 11. This model adds subject attributes, object attributes, and environmental attributes to the RBAC model and statically defines relationships such as subject-role, role-permission, and permission-object. In this access control model, the trust level of the subject, object, and environment is treated as attributes. Dynamic control of permissions is achieved through continuous trust evaluation.

The hybrid access control model comprises the following key components: *attribute selection*, *subject-role assignment and role-object assignment*, *dynamic trust evaluation*, and *permission filtering*.

1) *Attribute selection.* The hybrid access control model uses dynamic and static attribute sets for role assignment, trust evaluation, and permission filtering. The subject attribute (SATT) includes the subject's identity information, historical trust, purpose, and other information related to the subject type and state. The object attribute (OATT) includes data type, industry, security level, historical trust, and other information related to object type and security requirements. The environment attribute (EATT) includes information related to the current access behavior and the environment state, such as access time, network status, and the scenario of the access request.

TABLE II
APPLICATION SCENARIOS OF SECURE FL FRAMEWORK IN INDUSTRY

Industry	Application Scenario	Data	Client	Server
manufacturing	fault detection and diagnosis predictive maintenance production forecast quality control	production data quality data equipment data supply chain data	factory intelligent equipment	factory manufacturer company industry association
energy	energy production forecast energy consumption forecast energy optimal scheduling	energy production data energy consumption data environment data	smart meter energy service company micro-grid controller	energy scheduling center energy service company
transportation	vehicle condition monitoring driving behavior analysis traffic signal control navigation optimization	vehicle sensing data road condition data	vehicle road side unit	automobile manufacturer transport sector
healthcare	disease diagnosis medicine development chronic disease prediction gene association analysis health monitoring	medical data health data genetic data	home gateway hospital medical research institution	hospital medical research institution
chemical	reduce production risk improve production efficiency	production data equipment data	factory	scientific research institution factory
logistics	path planning and optimization improve transportation efficiency	road condition data cargo information vehicle sensing data	on board unit vehicle	logistics company

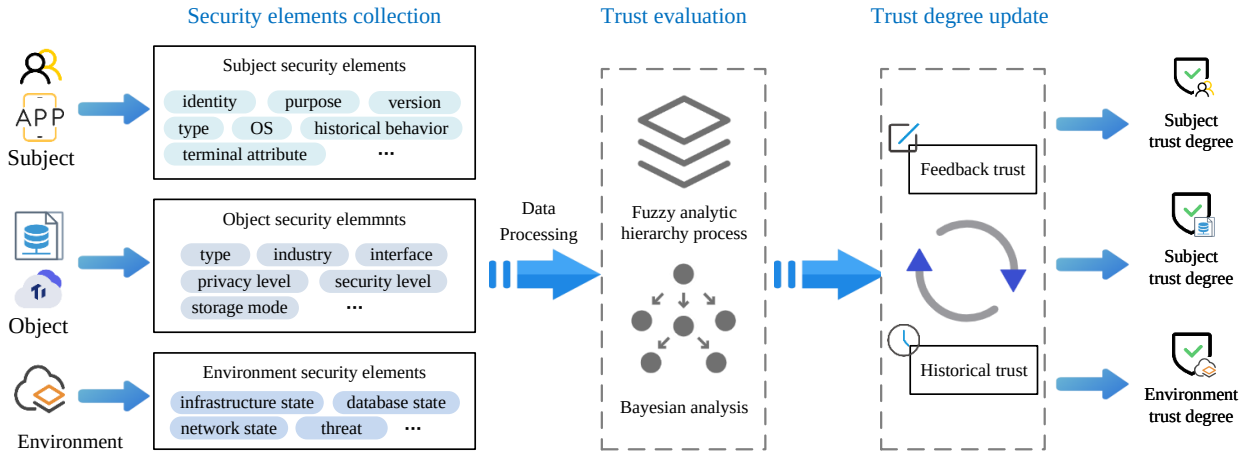

Fig. 10. Trust evaluation process.

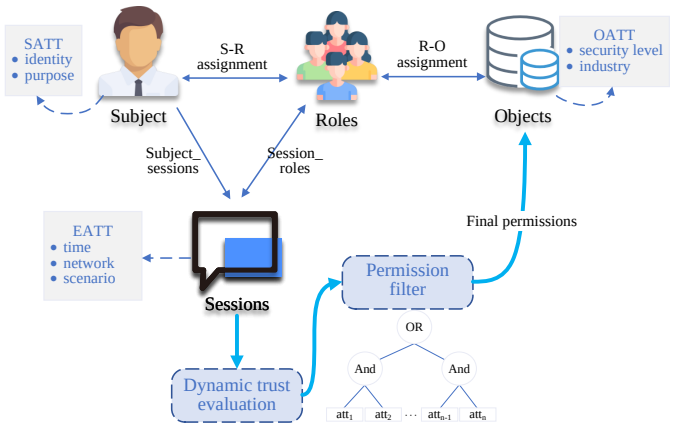

Fig. 11. Hybrid role-attribute access control model.

2) *Subject-role assignment and role-object assignment.* Subject-role assignment refers to assigning roles to subjects. A subject can only have the rights associated with a role after being assigned to that role. Role-object assignment refers to assigning objects (resources in the system) to roles. A user represented by a role can only access an object after the object is assigned to that role. The establishment of subject-role and role-object relations enables more flexible and convenient access control.
3) *Dynamic trust evaluation.* The process of trust evaluation, as shown in Fig. 10, is performed not only at the initiation of the access request but also continuously during the access. The results of each trust evaluation are used to update the historical trust degree of the subject, object, and environment.
4) *Permission filtering.* Permission filtering is the key to

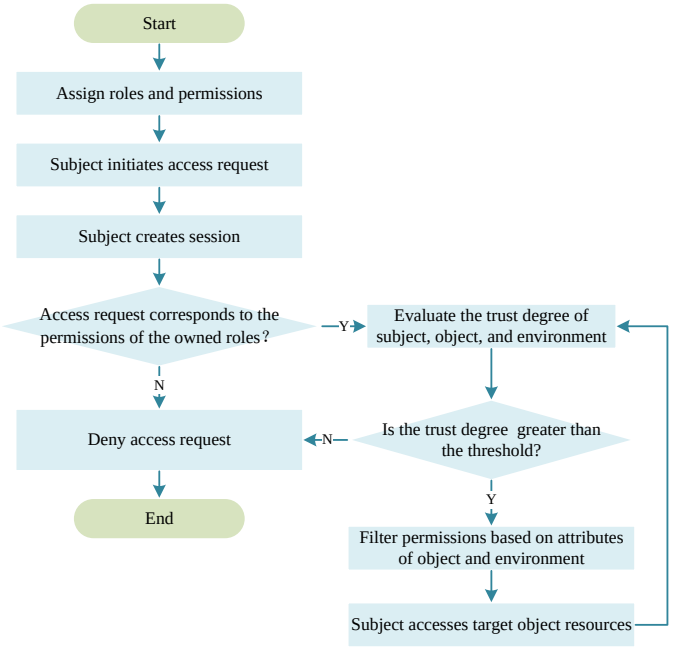

Fig. 12. The process of hybrid role-attribute access control.

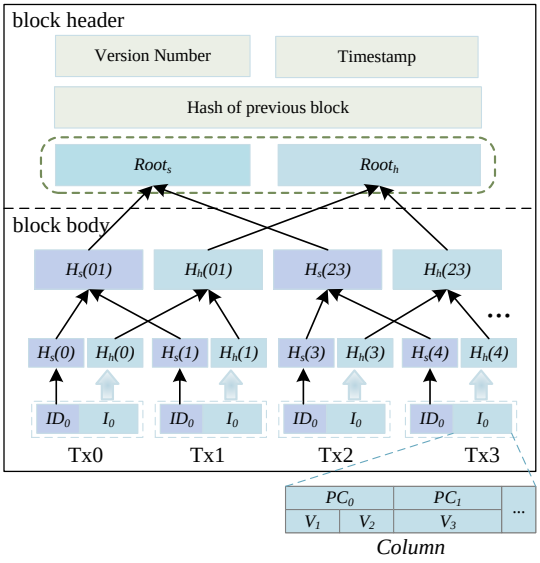

Fig. 13. Schematic diagram of scalable block structure.

achieving fine-grained access control. The restriction condition $cons(r, o, oatt, eatt)$ composed of each role r, object o, object attribute $oatt$, and environment attribute $eatt$ is used to delete the permission assigned to the subject. This process can be formalized as a binary tree, where the leaf node represents each attribute in the constraint, the intermediate node represents the and and or relation, and the value of the root node determines whether the constraint is satisfied. After permission filtering, the final permission set adheres to the principle of least privilege.

Fig. 12 shows the hybrid role-attribute access control process.

C. Blockchain Supporting Secure and Efficient IBDA

The block structure, consensus algorithms, and multiple smart contracts for secure FL and DS&T records are presented in this section.

1) Scalable Block Structure for IBD: Blockchain technology has numerous applications in industrial scenarios, including two significant applications closely related to IBDA: secure FL and DS&T. We first analyze the demands of these two applications for blockchain and then investigate suitable scalable block structures.

a) Blockchain Demand Analysis for IBD: Since the central server of a traditional FL system is susceptible to a single-point failure, using a blockchain system to perform aggregation tasks on distributed clients can enhance the efficiency and security of FL to some extent. However, it is difficult to store massive amounts of IBD directly in the blockchain ledger. A reasonable solution is to store original characteristic data off-chain, while only storing high-level data retrieval, model parameters, and other information in the blockchain.

In addition, data authentication, provenance, and auditing are challenges faced by IBD sharing and trading. Currently, data provenance and audit systems in the industry are heavily centralized and opaque, undermining their credibility. Blockchain system can use smart contracts to achieve standardized and automated DS&T recording, providing reliable data provenance and audit functions for IBDA platforms.

b) Scalable Block Structure for IBDA: The differences in the presentation methods of IBD result in varying tensor dimensions after data fusion and feature extraction. Additionally, data indexing methods also differ. As a result, it is essential to research a scalable block structure that can accommodate diverse business scenarios. This will allow blockchain ledgers to store various data and models in a unified data structure. A scalable block structure for IBDA is proposed as a reference, as shown in Fig. 13. The block header includes the block number, previous block hash value, current block hash value, and timestamp. It also contains two Merkle trees, $Root_h$ and $Root_s$. $Root_h$ stores the hash values of all data within the block, serving as the basis for block validation and consensus. And $Root_s$ is a Merkle tree composed of leaf nodes based on the identifiers of block data and model records. This structure can achieve more efficient heterogeneous data retrieval and enhance data availability.

Each record in the block body uses the $Key - Column$ model. Similar to the $Key - Value$ model, the Key serves as the unique entry point for searching for each data record. The $Column$ simulates the storage format of traditional tables through multi-layer mapping, providing high scalability for multi-group data. In our block structure, the Key field stores the identifier of the data and model, which is the leaf node of $Root_s$. The $Column$ field varies depending on content type. Table III shows $Column$ storage entries for key applications. Specifically, for tensor data after data fusion and feature extraction, it can be stored in the form of an n-dimensional array or use data pointers to point to offline storage locations to reduce storage overhead.

TABLE III
EXAMPLE OF *Column* DATA STRUCTURE

Data type	Storage items
data sharing records	provider, requestor, operator, operation time...
data trading records	consumer, provider, trading amount, trading time...
feature data (tensor form)	n-dimensional array
key model parameters	parameter type, parameter value...
decision results	decision time, decision content...
...	...

2) Consensus Algorithms for FL and DS&T: The blockchain system that supports FL must transmit key models and parameters with high data throughput. DS&T recording has high requirements for data reliability and supervision. As a result, it is necessary to research consensus algorithms that are tailored to meet the specific needs of different applications.

a) Efficient Hybrid Consensus Algorithm for FL: To meet the high-throughput and high-efficiency requirements of FL scenarios, a hybrid consensus mechanism based on PoS and pipelined Byzantine Fault Tolerance (BFT) can be adopted. The PoS consensus method is used to elect a specific committee responsible for sharding transactions in the network. Within the committee, the pipelined BFT consensus mechanism is run to generate blocks.

Committee election is based on the PoS consensus mechanism. In FL scenarios, the stake of a node can be measured based on the amount of data contributed, the quality of the uploaded model updates, and the computational resources consumed. Moreover, every node that completes model collaborative training receives an equity reward, increasing its weight for entering the committee in the next round election. On the other hand, a node will suffer severe penalties if it engages in poison attacks, fails to authenticate, or experiences significant delays. Since nodes within the committee can obtain accounting rights. This mechanism can motivate participants in FL to participate more actively and honestly in model collaborative training, thereby increasing their electoral weight.

The committee's internal consensus algorithm is based on the pipelined BFT mechanism, which is an improvement over the Practical Byzantine Fault Tolerance (PBFT) algorithm. With PBFT, each block must pass through three stages of voting and information interaction between nodes. However, the pipelined BFT mechanism processes voting for blocks in a parallel manner. As illustrated in Fig. 14, the block B_n proposal is confirmed if it receives more than 2/3 of the votes after one round of voting. In the next round, the block B_{n+1} proposal is included, along with the final confirmation vote for block B_n. If the votes obtained in this round exceed 2/3, block B_n is considered to have passed the *Commit* stage, and block B_{n+1} proposal advances to the *Prepare* stage. This parallel pipelining approach can reduce latency and increase throughput to meet FL requirements.

b) Reputation Proof Consensus Algorithm for DS&T: The consensus mechanism based on reputation proof can meet the high reliability requirements of DS&T records. Under this mechanism, only nodes with a credit value greater than

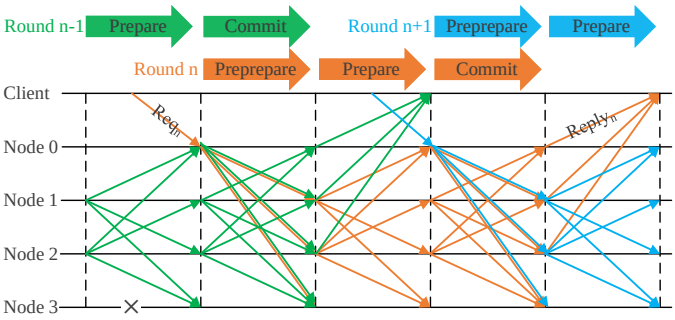

Fig. 14. The pipelined BFT mechanism.

the system's trust reference value are considered effective traceability nodes and can participate in the system's primary node election. This mechanism motivates nodes to share data and record transactions honestly and completely. If a node attempts to undermine the reliability of on-chain data by fabricating data or forking the blockchain, it will suffer a reputation penalty. Furthermore, this node will be in an untrusted position and face stronger supervision during data auditing and tracing. When the node can prove that its credit value in the current cycle is greater than the system's trust benchmark value α, it will enter the authority network through voting. The remaining consensus process is similar to Proof of Authority (PoA). The consensus mechanism mainly includes the following processes.

1) *Accounting node selection based on reputation proof.* At the start of each reputation cycle, each node generates an empty block linked to the previous block, from which a set of participants is derived. In each accounting round, N nodes are randomly selected from the participant set to serve as the basic equity representatives for the current round. These nodes must have reputation values greater than the reference value α.
2) *Block broadcast and ledger update.* Block broadcasting and ledger updates are basically consistent with the PoA consensus mechanism. The current accounting node collects, verifies, and broadcasts the data to be packaged within the current period, and each node verifies the identity of the accounting node. The absolute leadership of the accounting node in a single round effectively reduces the possibility of blockchain forks. The election of the accounting node is decentralized and regulated, thus mitigating the risk of intra-cycle centralization and preserving the decentralization of the overall consensus mechanism.
3) *Dynamic reputation adjustment.* At the end of each round of accounting, the reputation value needs to be dynamically adjusted, including rewards, punishments and centralized restrictions. The reputation value is calculated as follows:

$$T^i = \frac{1}{1 + e^{-\alpha(\sum_{j=0}^{n-1} P_j - \rho * \sum_{j=0}^{n-1} N_j)}}, \tag{14}$$

where, T^i is the reputation value of node i after this round of reputation evaluation; n is the total number of accounting rounds in this reputation cycle; $P_j, N_j \in$

$\{0, 1\}$ are normal voting and malicious voting in the j-th round, respectively. $\rho \in (0, 1)$ represents the penalty intensity for malicious voting behavior. Moreover, to prevent the reputation value from growing rapidly in the early stages of the blockchain system and leading to centralization, it is necessary to constrain the reputation value by the decentralized weight parameter η. The reputation value T^i should satisfy the following condition:

$$T^i < (1 + \eta)T^{i-1}. \tag{15}$$

3) Multiple Smart Contracts: On the blockchain system of IBDA scenarios, smart contracts can provide automated model preprocessing and transaction records and achieve trusted data interaction. For the two specific applications of FL and DS&T, it is necessary to study smart contracts that meet their security and efficiency requirements. We present multiple smart contracts that may be used on the IBDA platform, including but not limited to data storage smart contract (DSSC), data sharing smart contract (DHSC), data trading smart contract (DTSC), federated aggregation smart contract (FASC), and incentive smart contract (ISC).

1) *DSSC:* Using DSSC, data owners can submit data storage requests to blockchain nodes, and DSSC generates corresponding data blocks and data indexes for them and rewards nodes for contributing storage resources (see Alg. 1). The original data is encrypted, added with a digital signature by the data owner, and uploaded to the edge blockchain node, which audits the data using the DataAudit() function. The edge node periodically integrates the original data storage identities into the data block and broadcasts it to other edge nodes for verification. The storage proof process is performed by the ProofStorage() function in the blockchain cluster, and the storage resource reward is computed.
2) *DHSC:* DHSC provides data sharing audit, retrieval, recording, and reward functions (see Alg. 2). Firstly, the data user submits a data sharing request to the blockchain cluster, including the purpose, scope, and accuracy requirements of the requested data. Then, the blockchain cluster verifies the legitimacy of the data user's identity and the sharing request using the SharingValidate() function. If the verification is successful, the DataRetrieval() function is used to retrieve the data that satisfies the requirements in the blockchain and return the data index array to the data user. Finally, all the data sharing records are stored in the blockchain as blocks, and the data provides are rewarded according to the size of the shared data.
3) *DTSC:* DTSC implements the entire process of data trading (see Alg. 3). Firstly, the initiator of data trading submits a request that specifies the purpose, budget, scope, and requirements of the desired data. Then, the blockchain cluster verifies the legitimacy of the identity and the request of the initiator using the TradingValidate() function. If the verification is successful, the DataRetrieval() function searches for the data that meets the requirements in the blockchain and returns a list of data owners. The TradingNegotiate() function negotiates with the data owners and finalizes the data trading, returning the access address of the original data. Finally, the credentials of the data trading are stored in the blockchain as blocks.
4) *FASC:* FASC performs on-chain aggregation of model updates to generate an updated global model that can be accessed by all participants (see Alg. 4). During the global training round, FASC constantly retrieves local updates uploaded by clients in the blockchain and verifies their performance. When a sufficient number of updates are available, the model aggregation is triggered to generate a block of the new global model and publish it in the blockchain.
5) *ISC:* In FL, all clients voluntarily contribute their data and participate in model training. ISC incentivizes more data owners to join FL and provides data and computing resources (see Alg. 5). Firstly, the FL task initiator publishes its QoS requirements, such as accuracy requirements, number of devices required, dataset size required, training budget, and training time. After each client completes the training task within the training round, the ProofData() and ProofComputation() functions calculate the data rewards and computation rewards for the data owner, respectively.

We summarize in Table IV the advantages, disadvantages,

Algorithm 1 Data Storage Smart Contract (DSSC)

Input: raw_data: raw data provided by the data owner
Output: $data\_index$: data index in blockchain; $reward\_storage$: reward for the node providing storage resources

$sig \leftarrow$Signature(raw_data);
$encrypt_data \leftarrow$Encypt(raw_data);
Upload($encrypt\_data$, sig);
DataAudit($encrypt\_data$, sig);
$data_index \leftarrow$GeneBlock($encrypt\_data$);
$reward_storage \leftarrow$ProofStorage();
return $data\_index$, $reward\_storage$;

Algorithm 2 Data Sharing Smart Contract (DHSC)

Input: $request$: data sharing request from the data user
Output: $reward\_sharing$: reward for the data provider; $data\_indexs$: data indexs of all shared data in blockchain

$val \leftarrow$SharingValidate($request$);
if val **then**
    $data_indexs \leftarrow$DataRetrieval($request$);
    RecordSharing();
    $reward_sharing \leftarrow$ProofSharing();
end if
return $data\_indexs$, $reward\_sharing$;

Algorithm 3 Data Trading Smart Contract (DTSC)

Input: $request$: data trading request from the data user
Output: $data\_address$: access address of raw data

$val \leftarrow$TradingValidate($request$);
if val **then**
    $data_owners \leftarrow$DataRetrieval($request$);
    $data_address \leftarrow$TradingNegotiate($data_owners$, $request$)
    RecordTrading();
end if
return $data\_address$;

Algorithm 4 Federated Aggregate Smart Contract (FASC)

Input: $global\_round$: global traing rounds
Output: $global\_model$: new global model after aggregation

while $round \leq gloabl\_round$ **do**
  $local_update \leftarrow$QueryLocalUpdate();
  VerifyUpdate($local\_update$);
  $local_updats \leftarrow$Append($local\_update$);
  $global_model \leftarrow$Aggregate($local\_updates$);
  $round++$;
end while
return $global\_model$;

Algorithm 5 Incentive Smart Contract (ISC)

Input: qos_req: QoS requirements issued by the initiator; $global\_round$: global training rounds
Output: $reward\_data$: data reward for data owners; $reward\_comput$: computing power reward for data owners

$clients \leftarrow$Release(qos_req);
while $round \leq gloabl\_round$ **do**
  for $client$ in $clients$ **do**
    client downloads the global model, trains local model, and uploads local model update;
    $reward_data \leftarrow$ProofData();
    $reward_comput \leftarrow$ProofComputation();
  end for
  $round++$;
end while
return $reward\_data$, $reward\_comput$;

and applicability of the solutions for the three key technologies discussed above, along with their specific objectives for addressing the three challenges mentioned in Section III.

VI. Conclusions

As the construction of the Industrial Internet progresses, data-driven innovative application modes are gradually being explored. IBD provides significant value at the enterprise, social, and national levels through DS&T. However, there are three major challenges that must be addressed immediately: existing big data analytics methods cannot meet the new characteristics of IBD; DS&T raise privacy and security concerns; and the Industrial Internet environment lacks mutual trust. This paper proposed a research framework for privacy-preserving and secure IBDA and elaborated on the research proposals and potential technologies from four perspectives: platform architecture, data fusion and analytics methods, privacy and security protection methods, and blockchain supporting IBDA.

- In terms of platform architecture, we proposed a function model, a security architecture, and a system architecture for the IBDA platform. The platform can support secure access and storage of multi-source heterogeneous IBD, multi-modal fusion and analytics, and privacy-preserving and efficient sharing and trading. These provide a reference for constructing a privacy-preserving and secure IBDA platform.
- In terms of data fusion and analytics, we first suggested using a multi-scale heterogeneous graph model to implement IBD joint representation and reveal the inherent multi-granularity features of multi-modal IBD. We then proposed using a prototype-based multi-modal intelligence fusion analytics approach to improve domain adaptation and interpretability. Finally, we discussed distributed elastic computing methods for data, task, and model extension.
- In terms of data privacy and security protection, we first recommended identifying the sensitive attributes of IBD to support hierarchical classification of privacy protection. We then proposed using a blockchain and TEE enhanced secure FL framework for distributed, privacy-preserving IBD modeling. Finally, referring to the concept of zero-trust, we suggested using dynamic trust evaluation and a role-attribute hybrid access control model to realize secure data access in the Industrial Internet.
- In terms of blockchain, we first designed a scalable block structure to meet the various data storage and retrieval requirements of IBD scenarios. We then provided two reference consensus algorithms for secure FL and DS&T. Finally, we proposed multiple smart contracts to meet the security and efficiency requirements of various IBDA platform applications and discussed the content of key smart contracts.

To the best of our knowledge, this paper is the first to propose a research framework for privacy-preserving and secure IBDA, addressing three challenges to secure and efficient IBDA. This work has guiding significance for IBDA and platform construction and can benefit IBDA researchers and industry practitioners by providing clear guidance for constructing IBDA platforms in various industries, particularly in terms of system architecture and technical solutions.

References

[1] G. Tang, Z. Feng, D. Li, and X. Ai, "Summary and prospect of industrial internet: based on bibliometric analysis," *Comput. Integr. Manuf. Syst.*, pp. 1–21, 2023.
[2] "Global industrial internet innovation and development report," 2022. [Online]. Available: https://www.china-aii.com/filedownload/680549
[3] J. Wang, P. Zheng, Y. Lv, J. Bao, and J. Zhang, "Fog-ibdis: industrial big data integration and sharing with fog computing for manufacturing systems," *Engineering*, vol. 5, no. 4, pp. 662–670, 2019.
[4] H. Chen, R. Wang, X. Liu, Y. Du, and Y. Yang, "Monitoring the enterprise carbon emissions using electricity big data: A case study of beijing," *J. Clean Prod.*, vol. 396, p. 136427, 2023.
[5] K. Demertzis, L. Iliadis, and I. Bougoudis, "Gryphon: a semi-supervised anomaly detection system based on one-class evolving spiking neural network," *Neural Comput. Appl.*, vol. 32, pp. 4303–4314, 2020.
[6] "Trusted industrial data space system architecture 1.0," 2022. [Online]. Available: http://www.aii-alliance.org/uploads/1/20220125/68c2389362c6f6005711ac4e68e40425.pdf
[7] "2022 data compromise cost report," 2022. [Online]. Available: https://www.ibm.com/downloads/cas/A48NDEYW
[8] A. C. Ikegwu, H. F. Nweke, C. V. Anikwe, U. R. Alo, and O. R. Okonkwo, "Big data analytics for data-driven

industry: a review of data sources, tools, challenges, solutions, and research directions," *Cluster Comput.*, vol. 25, no. 5, pp. 3343–3387, 2022.

[9] Z. Lv, H. Song, P. Basanta-Val, A. Steed, and M. Jo, "Next-generation big data analytics: State of the art, challenges, and future research topics," *IEEE Trans. Ind. Inform.*, vol. 13, no. 4, pp. 1891–1899, 2017.

[10] S. K. Jagatheesaperumal, M. Rahouti, K. Ahmad, A. Al-Fuqaha, and M. Guizani, "The duo of artificial intelligence and big data for industry 4.0: Applications, techniques, challenges, and future research directions," *IEEE Internet Things J.*, vol. 9, no. 15, pp. 12 861–12 885, 2021.

[11] M. Khan, X. Wu, X. Xu, and W. Dou, "Big data challenges and opportunities in the hype of industry 4.0," in *IEEE Int Conf. Commun.* IEEE, 2017, pp. 1–6.

[12] M. Javaid, A. Haleem, R. P. Singh, and R. Suman, "Significant applications of big data in industry 4.0," *J. Ind. Integr. Manag.*, vol. 6, no. 04, pp. 429–447, 2021.

[13] J.-Q. Li, F. R. Yu, G. Deng, C. Luo, Z. Ming, and Q. Yan, "Industrial internet: A survey on the enabling technologies, applications, and challenges," *IEEE Commun. Surv. Tutor.*, vol. 19, no. 3, pp. 1504–1526, 2017.

[14] R. Bonnard, M. D. S. Arantes, R. Lorbieski, K. M. M. Vieira, and M. C. Nunes, "Big data/analytics platform for industry 4.0 implementation in advanced manufacturing context," *Int. J. Adv. Manuf. Technol.*, vol. 117, no. 5-6, pp. 1959–1973, 2021.

[15] M. Y. Santos, J. Oliveira e Sá, C. Costa, J. Galvão, C. Andrade, B. Martinho, F. V. Lima, and E. Costa, "A big data analytics architecture for industry 4.0," in *Adv. Intell. Sys. Comput.* Springer, 2017, pp. 175–184.

[16] C. Li, Y. Chen, and Y. Shang, "A review of industrial big data for decision making in intelligent manufacturing," *Eng. Sci. Technol.*, vol. 29, p. 101021, 2022.

[17] J. Wang, C. Xu, J. Zhang, and R. Zhong, "Big data analytics for intelligent manufacturing systems: A review," *J. Manuf. Syst.*, vol. 62, pp. 738–752, 2022.

[18] J. Zhou, X. Yao, and J. Zhang, "Big data in wisdom manufacturing for industry 4.0," in *Int. Conf. Enterp. Syst.* IEEE, 2017, pp. 107–112.

[19] S. Sarker, M. S. Arefin, M. Kowsher, T. Bhuiyan, P. K. Dhar, and O.-J. Kwon, "A comprehensive review on big data for industries: Challenges and opportunities," *IEEE Access*, 2022.

[20] K. Zhou, C. Fu, and S. Yang, "Big data driven smart energy management: From big data to big insights," *Renew. Sust. Energ. Rev.*, vol. 56, pp. 215–225, 2016.

[21] E. Hossain, I. Khan, F. Un-Noor, S. S. Sikander, and M. S. H. Sunny, "Application of big data and machine learning in smart grid, and associated security concerns: A review," *Ieee Access*, vol. 7, pp. 13 960–13 988, 2019.

[22] T. Nguyen, R. G. Gosine, and P. Warrian, "A systematic review of big data analytics for oil and gas industry 4.0," *IEEE Access*, vol. 8, pp. 61 183–61 201, 2020.

[23] H. Eissa, "Unleashing industry 4.0 opportunities: Big data analytics in the midstream oil & gas sector," in *Int. Pet. Technol. Conf.* IPTC, 2020, p. D033S076R002.

[24] S. Kaffash, A. T. Nguyen, and J. Zhu, "Big data algorithms and applications in intelligent transportation system: A review and bibliometric analysis," *Int. J. Prod. Econ.*, vol. 231, p. 107868, 2021.

[25] M. Bilal, L. O. Oyedele, J. Qadir, K. Munir, S. O. Ajayi, O. O. Akinade, H. A. Owolabi, H. A. Alaka, and M. Pasha, "Big data in the construction industry: A review of present status, opportunities, and future trends," *Adv. Eng. Inform.*, vol. 30, no. 3, pp. 500–521, 2016.

[26] F. Li, Y. Laili, X. Chen, Y. Lou, C. Wang, H. Yang, X. Gao, and H. Han, "Towards big data driven construction industry," *Journal of Industrial Information Integration*, p. 100483, 2023.

[27] M. Karatas, L. Eriskin, M. Deveci, D. Pamucar, and H. Garg, "Big data for healthcare industry 4.0: Applications, challenges and future perspectives," *Expert Syst. Appl.*, vol. 200, p. 116912, 2022.

[28] S. Sankar, B. Shadaksharappa, N. G. Kumar, and K. Padmanaban, "Big data analytics-based intelligent logistic system," in *Int. Conf. Electron. Sustain. Commun. Syst.* IEEE, 2023, pp. 1190–1195.

[29] J. Koo, G. Kang, and Y.-G. Kim, "Security and privacy in big data life cycle: a survey and open challenges," *Sustainability*, vol. 12, no. 24, p. 10571, 2020.

[30] D. Demirol, R. Das, and D. Hanbay, "A key review on security and privacy of big data: issues, challenges, and future research directions," *Signal Image Video Process.*, pp. 1–9, 2022.

[31] M. I. Pramanik, R. Y. Lau, M. S. Hossain, M. M. Rahoman, S. K. Debnath, M. G. Rashed, and M. Z. Uddin, "Privacy preserving big data analytics: A critical analysis of state-of-the-art," *Wiley Interdiscip. Rev.-Data Mining Knowl. Discov.*, vol. 11, no. 1, p. e1387, 2021.

[32] A. Amaithi Rajan and V. V, "Systematic survey: Secure and privacy-preserving big data analytics in cloud," *J. Comput. Inf. Syst.*, pp. 1–21, 2023.

[33] N. Deepa, Q.-V. Pham, D. C. Nguyen, S. Bhattacharya, B. Prabadevi, T. R. Gadekallu, P. K. R. Maddikunta, F. Fang, and P. N. Pathirana, "A survey on blockchain for big data: approaches, opportunities, and future directions," *Futur. Gener. Comp. Syst.*, 2022.

[34] M. Amiri-Zarandi, R. A. Dara, E. Duncan, and E. D. Fraser, "Big data privacy in smart farming: a review," *Sustainability*, vol. 14, no. 15, p. 9120, 2022.

[35] W. N. Price and I. G. Cohen, "Privacy in the age of medical big data," *Nat. Med.*, vol. 25, no. 1, pp. 37–43, 2019.

[36] Z. Guo, Y. Luo, Z. Cai, and T. Zheng, "Overview of privacy protection technology of big data in healthcare," *J. Frontiers Comput. Sci. Technolo.*, vol. 15, no. 3, p. 389, 2021.

[37] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," *arXiv preprint arXiv:1707.07250*, 2017.

[38] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang,

A. Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," *arXiv preprint arXiv:1806.00064*, 2018.

[39] J. He, C. Zhang, X. Li, and D. Zhang, "Survey of research on multimodal fusion technology for deep learning," *Computer Engineering*, vol. 46, no. 5, pp. 1–11, 2020.

[40] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L.-P. Morency, "Memory fusion network for multi-view sequential learning," in *AAAI Conf. Artif. Intell.*, vol. 32, no. 1, 2018.

[41] L. Mou, C. Zhou, P. Zhao, B. Nakisa, M. N. Rastgoo, R. Jain, and W. Gao, "Driver stress detection via multimodal fusion using attention-based cnn-lstm," *Expert Syst. Appl.*, vol. 173, p. 114693, 2021.

[42] J. Lv, K. Liu, and S. He, "Differentiated learning for multi-modal domain adaptation," in *Proc. ACM Int. Conf. Multimed.*, 2021, pp. 1322–1330.

[43] S. Ma, K. Feng, J. Li, H. Wang, G. Cong, and J. Huai, "Proxies for shortest path and distance queries," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 7, pp. 1835–1850, 2016.

[44] L. Duan, C. Aggarwal, S. Ma, R. Hu, and J. Huai, "Scaling up link prediction with ensembles," in *ACM Int. Conf. Web Search Data Min.*, 2016, pp. 367–376.

[45] A. P. Iyer, Z. Liu, X. Jin, S. Venkataraman, V. Braverman, and I. Stoica, "Asap: Fast, approximate graph pattern mining at scale." in *OSDI*, vol. 18, 2018, pp. 745–761.

[46] X. Lin, S. Ma, H. Zhang, T. Wo, and J. Huai, "One-pass error bounded trajectory simplification," *arXiv preprint arXiv:1702.05597*, 2017.

[47] S. Ma, R. Hu, L. Wang, X. Lin, and J. Huai, "Fast computation of dense temporal subgraphs," in *Proc. Int. Conf. Data Eng.* IEEE, 2017, pp. 361–372.

[48] S. Mittal, "A survey of techniques for approximate computing," *ACM Comput. Surv.*, vol. 48, no. 4, pp. 1–33, 2016.

[49] G. Zervakis, "On accelerating data analytics: An introduction to the approximate computing technique," *IoT for Smart Grids: Design Challenges and Paradigms*, pp. 163–180, 2019.

[50] T. Liu, W. Chen, T. Wang, and F. Gao, *Distributed Machine Learning*. China Machine Press, 2018.

[51] J. Dean and S. Ghemawat, "Mapreduce: simplified data processing on large clusters," *Commun. ACM*, vol. 51, no. 1, pp. 107–113, 2008.

[52] M. Assefi, E. Behravesh, G. Liu, and A. P. Tafti, "Big data machine learning using apache spark mllib," in *IEEE Int. Conf. Big Data*. IEEE, 2017, pp. 3492–3498.

[53] A. Beygelzimer, H. Daumé, J. Langford, and P. Mineiro, "Learning reductions that really work," *Proc. IEEE*, vol. 104, no. 1, pp. 136–147, 2015.

[54] E. P. Xing, Q. Ho, W. Dai, J.-K. Kim, J. Wei, S. Lee, X. Zheng, P. Xie, A. Kumar, and Y. Yu, "Petuum: A new platform for distributed machine learning on big data," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2015, pp. 1335–1344.

[55] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, M. Mao, M. Ranzato, A. Senior, P. Tucker, K. Yang *et al.*, "Large scale distributed deep networks," *Adv. neural inf. proces. syst.*, vol. 25, 2012.

[56] J. Zhu, Z. Ge, and Z. Song, "Distributed parallel pca for modeling and monitoring of large-scale plant-wide processes with big data," *IEEE Trans. Ind. Inform.*, vol. 13, no. 4, pp. 1877–1885, 2017.

[57] D. Xia, M. Zhang, X. Yan, Y. Bai, Y. Zheng, Y. Li, and H. Li, "A distributed wnd-lstm model on mapreduce for short-term traffic flow prediction," *Neural Comput. Appl.*, vol. 33, pp. 2393–2410, 2021.

[58] B.-C. Liu, A. Binaykia, P.-C. Chang, M. K. Tiwari, and C.-C. Tsao, "Urban air quality forecasting based on multi-dimensional collaborative support vector regression (svr): A case study of beijing-tianjin-shijiazhuang," *PloS one*, vol. 12, no. 7, p. e0179763, 2017.

[59] R. Fezai, K. Dhibi, M. Mansouri, M. Trabelsi, M. Hajji, K. Bouzrara, H. Nounou, and M. Nounou, "Effective random forest-based fault detection and diagnosis for wind energy conversion systems," *IEEE Sens. J.*, vol. 21, no. 5, pp. 6914–6921, 2020.

[60] Q. Xu, Z. Fan, W. Jia, and C. Jiang, "Quantile regression neural network-based fault detection scheme for wind turbines with application to monitoring a bearing," *Wind Energy*, vol. 22, no. 10, pp. 1390–1401, 2019.

[61] M. Samie Tootooni, A. Dsouza, R. Donovan, P. K. Rao, Z. Kong, and P. Borgesen, "Classifying the dimensional variation in additive manufactured parts from laser-scanned three-dimensional point cloud data using machine learning approaches," *J. Manuf. Sci. Eng.-Trans. ASME*, vol. 139, no. 9, p. 091005, 2017.

[62] M. Liukkonen and Y. Hiltunen, "Recognition of systematic spatial patterns in silicon wafers based on som and k-means," *IFAC-PapersOnLine*, vol. 51, no. 2, pp. 439–444, 2018.

[63] K. Nakata, R. Orihara, Y. Mizuoka, and K. Takagi, "A comprehensive big-data-based monitoring system for yield enhancement in semiconductor manufacturing," *IEEE Trans. Semicond. Manuf.*, vol. 30, no. 4, pp. 339–344, 2017.

[64] M. Onel, B. Beykal, K. Ferguson, W. A. Chiu, T. J. McDonald, L. Zhou, J. S. House, F. A. Wright, D. A. Sheen, I. Rusyn *et al.*, "Grouping of complex substances using analytical chemistry data: A framework for quantitative evaluation and visualization," *PloS one*, vol. 14, no. 10, p. e0223517, 2019.

[65] J. Hou, Y. Wu, A. S. Ahmad, H. Gong, and L. Liu, "A novel rolling bearing fault diagnosis method based on adaptive feature selection and clustering," *Ieee Access*, vol. 9, pp. 99 756–99 767, 2021.

[66] C. H. Jin, H. J. Na, M. Piao, G. Pok, and K. H. Ryu, "A novel dbscan-based defect pattern detection and classification framework for wafer bin map," *IEEE Trans. Semicond. Manuf.*, vol. 32, no. 3, pp. 286–292, 2019.

[67] A. Telikani and A. Shahbahrami, "Data sanitization in association rule mining: An analytical review," *Expert*

Syst. Appl., vol. 96, pp. 406–426, 2018.

[68] B. Jiang, J. Li, G. Yue, and H. Song, "Differential privacy for industrial internet of things: Opportunities, applications, and challenges," *IEEE Internet Things J.*, vol. 8, no. 13, pp. 10 430–10 451, 2021.

[69] M. Du, K. Wang, Z. Xia, and Y. Zhang, "Differential privacy preserving of training model in wireless big data with edge computing," *IEEE Trans. Big Data*, vol. 6, no. 2, pp. 283–295, 2018.

[70] P. Zhou, K. Wang, J. Xu, and D. Wu, "Differentially-private and trustworthy online social multimedia big data retrieval in edge computing," *IEEE Trans. Multimedia*, vol. 21, no. 3, pp. 539–554, 2018.

[71] S. Li, S. Zhao, G. Min, L. Qi, and G. Liu, "Lightweight privacy-preserving scheme using homomorphic encryption in industrial internet of things," *IEEE Internet Things J.*, vol. 9, no. 16, pp. 14 542–14 550, 2021.

[72] R. Lu, "A new communication-efficient privacy-preserving range query scheme in fog-enhanced iot," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 2497–2505, 2018.

[73] M. Keller, V. Pastro, and D. Rotaru, "Overdrive: making spdz great again," in *Annu. Int. Conf. Theory Appl. Cryptographic Techn.* Springer, 2018, pp. 158–189.

[74] D. Demmler, T. Schneider, and M. Zohner, "Aby-a framework for efficient mixed-protocol secure two-party computation." in *Annu. Netw. Distrib. Syst. Secur. Symp.*, 2015.

[75] M. Hastings, B. Hemenway, D. Noble, and S. Zdancewic, "Sok: General purpose compilers for secure multi-party computation," in *Proc. IEEE Symp. Secur. Privacy*. IEEE, 2019, pp. 1220–1237.

[76] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. Int. Conf. Artif. Intell. Stat.* PMLR, 2017, pp. 1273–1282.

[77] C. Zhou, D. Chen, S. Wang, A. Fu, and Y. Gao, "Research and challenge of distributed deep learning privacy and security attack," *J. Comput. Res. Develop.*, vol. 58, no. 5, pp. 927–943, 2021.

[78] K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal, and K. Seth, "Practical secure aggregation for privacy-preserving machine learning," in *Proc. ACM Conf. Computer Commun. Secur.*, 2017, pp. 1175–1191.

[79] R. C. Geyer, T. Klein, and M. Nabi, "Differentially private federated learning: A client level perspective," *arXiv preprint arXiv:1712.07557*, 2017.

[80] F. Tramer and D. Boneh, "Slalom: Fast, verifiable and private execution of neural networks in trusted hardware," *arXiv preprint arXiv:1806.03287*, 2018.

[81] Y. Lu, X. Huang, Y. Dai, S. Maharjan, and Y. Zhang, "Blockchain and federated learning for privacy-preserved data sharing in industrial iot," *IEEE Trans. Ind. Inform.*, vol. 16, no. 6, pp. 4177–4186, 2019.

[82] A. Triastcyn and B. Faltings, "Bayesian differential privacy for machine learning," in *Int. Conf. Machin. Learn.* PMLR, 2020, pp. 9583–9592.

[83] L. Yu, L. Liu, C. Pu, M. E. Gursoy, and S. Truex, "Differentially private model publishing for deep learning," in *Proc. IEEE Symp. Secur. Privacy*. IEEE, 2019, pp. 332–349.

[84] H. Kim, J. Park, M. Bennis, and S.-L. Kim, "Blockchained on-device federated learning," *IEEE Commun. Lett.*, vol. 24, no. 6, pp. 1279–1283, 2019.

[85] Z. Li, X. Gui, Y. Gu, X. Li, H. Dai, and X. Zhang, "Survey on homomorphic encryption algorithm and its application in the privacy-preserving for cloud computing," *J. Softw.*, vol. 29, no. 7, pp. 1830–1851, 2017.

[86] N. Jayapandian and A. Md Zubair Rahman, "Secure deduplication for cloud storage using interactive message-locked encryption with convergent encryption, to reduce storage space," *Braz. Arch. Biol. Technol.*, vol. 61, 2018.

[87] H. Deng, F. Song, L. Fu, L. Ou, H. Yin, Y. Gao, and Z. Qin, "A survey of data security and privacy preserving in cloud computing," *J. Hunan Univ.(Natural Sciences)*, vol. 49, no. 4, pp. 1–10, 2022.

[88] H. Zhong, W. Zhu, Y. Xu, and J. Cui, "Multi-authority attribute-based encryption access control scheme with policy hidden for cloud storage," *Soft Comput.*, vol. 22, pp. 243–251, 2018.

[89] Y. Tao, Z. Lei, and P. Ruxiang, "Fine-grained big data security method based on zero trust model," in *Proc. Int. Conf. Parallel Distrib. Syst.* IEEE, 2018, pp. 1040–1045.

[90] Z. Zaheer, H. Chang, S. Mukherjee, and J. Van der Merwe, "eztrust: Network-independent zero-trust perimeterization for microservices," in *Proc. ACM Symp. SDN Res.*, 2019, pp. 49–61.

[91] B. Chen, S. Qiao, J. Zhao, D. Liu, X. Shi, M. Lyu, H. Chen, H. Lu, and Y. Zhai, "A security awareness and protection system for 5g smart healthcare based on zero-trust architecture," *IEEE Internet Things J.*, vol. 8, no. 13, pp. 10 248–10 263, 2020.

[92] Y. Qu, S. R. Pokhrel, S. Garg, L. Gao, and Y. Xiang, "A blockchained federated learning framework for cognitive computing in industry 4.0 networks," *IEEE Trans. Ind. Inform.*, vol. 17, no. 4, pp. 2964–2973, 2020.

[93] D. Unal, M. Hammoudeh, M. A. Khan, A. Abuarqoub, G. Epiphaniou, and R. Hamila, "Integration of federated machine learning and blockchain for the provision of secure big data analytics for internet of things," *Comput. Secur.*, vol. 109, p. 102393, 2021.

[94] B. Jia, X. Zhang, J. Liu, Y. Zhang, K. Huang, and Y. Liang, "Blockchain-enabled federated learning data protection aggregation scheme with differential privacy and homomorphic encryption in iiot," *IEEE Trans. Ind. Inform.*, vol. 18, no. 6, pp. 4049–4058, 2021.

[95] S. Gao, L. Yuan, J. Zhu, X. Ma, R. Zhang, and J. Ma, "A blockchain-based privacy-preserving asynchronous federated learning," *Scientia Sinica Informationis*, vol. 51, pp. 1755–1774, 2021.

[96] M. H. ur Rehman, K. Salah, E. Damiani, and D. Svetinovic, "Towards blockchain-based reputation-aware federated learning," in *IEEE INFOCOM - IEEE*

Conf. Comput. Commun. Workshops. IEEE, 2020, pp. 183–188.

[97] H. Moudoud, S. Cherkaoui, and L. Khoukhi, "Towards a secure and reliable federated learning using blockchain," in *IEEE Glob. Commun. Conf.* IEEE, 2021, pp. 01–06.

[98] Y. J. Kim and C. S. Hong, "Blockchain-based node-aware dynamic weighting methods for improving federated learning performance," in *Asia-Pacific Netw. Oper. Manag. Symp.: Manag. Cyber-Physical World*. IEEE, 2019, pp. 1–4.

[99] S. R. Pokhrel, "Federated learning meets blockchain at 6g edge: A drone-assisted networking for disaster response," in *Proc. ACM MobiCom Workshop Drone Assist. Wirel. Commun. 5G Beyond*, 2020, pp. 49–54.

[100] Y. Qu, S. R. Pokhrel, S. Garg, L. Gao, and Y. Xiang, "A blockchained federated learning framework for cognitive computing in industry 4.0 networks," *IEEE Trans. Ind. Inform.*, vol. 17, no. 4, pp. 2964–2973, 2020.

[101] S. R. Pokhrel and J. Choi, "Federated learning with blockchain for autonomous vehicles: Analysis and design challenges," *IEEE Trans. Commun.*, vol. 68, no. 8, pp. 4734–4746, 2020.

[102] Y. Lu, X. Huang, K. Zhang, S. Maharjan, and Y. Zhang, "Low-latency federated learning and blockchain for edge association in digital twin empowered 6g networks," *IEEE Trans. Ind. Inform.*, vol. 17, no. 7, pp. 5098–5107, 2020.

[103] Y. Li, C. Chen, N. Liu, H. Huang, Z. Zheng, and Q. Yan, "A blockchain-based decentralized federated learning framework with committee consensus," *IEEE Netw.*, vol. 35, no. 1, pp. 234–241, 2020.

[104] Y. Lu, X. Huang, Y. Dai, S. Maharjan, and Y. Zhang, "Blockchain and federated learning for privacy-preserved data sharing in industrial iot," *IEEE Trans. Ind. Inform.*, vol. 16, no. 6, pp. 4177–4186, 2019.

[105] M. H. ur Rehman, A. M. Dirir, K. Salah, E. Damiani, and D. Svetinovic, "Trustfed: A framework for fair and trustworthy cross-device federated learning in iiot," *IEEE Trans. Ind. Inform.*, vol. 17, no. 12, pp. 8485–8494, 2021.

[106] V. Mugunthan, R. Rahman, and L. Kagal, "Blockflow: decentralized, privacy-preserving, and accountable federated machine learning," in *Blockchain Appl.: 3rd Int. Congr.* Springer, 2022, pp. 233–242.

[107] W. Zhang, Q. Lu, Q. Yu, Z. Li, Y. Liu, S. K. Lo, S. Chen, X. Xu, and L. Zhu, "Blockchain-based federated learning for device failure detection in industrial iot," *IEEE Internet Things J.*, vol. 8, no. 7, pp. 5926–5937, 2020.

[108] I. Martinez, S. Francis, and A. S. Hafid, "Record and reward federated learning contributions with blockchain," in *Int. Conf. Cyber-Enabled Distrib. Comput. Knowl. Discov.* IEEE, 2019, pp. 50–57.

[109] H. Lee and J. Kim, "Trends in blockchain and federated learning for data sharing in distributed platforms," in *Int. Conf. Ubiquitous Future Netw.* IEEE, 2021, pp. 430–433.

[110] Y. Zhao, Y. Yu, Y. Li, G. Han, and X. Du, "Machine learning based privacy-preserving fair data trading in big data market," *Inf. Sci.*, vol. 478, pp. 449–460, 2019.

[111] W. Dai, C. Dai, K.-K. R. Choo, C. Cui, D. Zou, and H. Jin, "Sdte: A secure blockchain-based data trading ecosystem," *IEEE Trans. Inf. Forensic Secur.*, vol. 15, pp. 725–737, 2019.

[112] J. Cui, F. Ouyang, Z. Ying, L. Wei, and H. Zhong, "Secure and efficient data sharing among vehicles based on consortium blockchain," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 8857–8867, 2021.

[113] D. Hu, Y. Li, L. Pan, M. Li, and S. Zheng, "A blockchain-based trading system for big data," *Comput. Netw.*, vol. 191, p. 107994, 2021.

[114] S. Zheng, L. Pan, D. Hu, M. Li, and Y. Fan, "A blockchain-based trading platform for big data," in *IEEE INFOCOM - IEEE Conf. Comput. Commun. Workshops*. IEEE, 2020, pp. 991–996.

[115] Y. Jiang, Y. Zhong, and X. Ge, "Smart contract-based data commodity transactions for industrial internet of things," *IEEE Access*, vol. 7, pp. 180 856–180 866, 2019.

[116] J. Kang, R. Yu, X. Huang, M. Wu, S. Maharjan, S. Xie, and Y. Zhang, "Blockchain for secure and efficient data sharing in vehicular edge computing and networks," *IEEE Internet Things J.*, vol. 6, no. 3, pp. 4660–4670, 2018.

[117] Y. Muchhala, H. Singhania, S. Sheth, and K. Devadkar, "Enabling mapreduce based parallel computation in smart contracts," in *Proc. Int. Conf. Inven. Comput. Technol.* IEEE, 2021, pp. 537–543.

[118] M. Cheng, Q. Li, J. Lv, W. Liu, and J. Wang, "Multi-scale lstm model for bgp anomaly classification," *IEEE Trans. Serv. Comput.*, vol. 14, no. 3, pp. 765–778, 2018.

[119] J. Lv, Q. Sun, Q. Li, and L. Moreira-Matias, "Multi-scale and multi-scope convolutional neural networks for destination prediction of trajectories," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 8, pp. 3184–3195, 2019.

[120] W. Mo and J. Lv, "Cascaded hierarchical context-aware vehicle re-identification," in *Proc Int Jt Conf Neural Networks*. IEEE, 2021, pp. 1–8.

[121] J. Lv, J. Liang, and Z. Yang, "Hge2med: Heterogeneous graph embedding for multi-domain event detection," in *Proc. Int. Conf. Tools Artif. Intell.* IEEE, 2020, pp. 1036–1043.

[122] J. Lv, J. Zhong, J. Liang, and Z. Yang, "Ace: Ant colony based multi-level network embedding for hierarchical graph representation learning," *IEEE Access*, vol. 7, pp. 73 970–73 982, 2019.

[123] A. Van Looveren and J. Klaise, "Interpretable counterfactual explanations guided by prototypes," in *Lect. Notes Comput. Sci.* Springer, 2021, pp. 650–665.

[124] F.-L. Fan, J. Xiong, M. Li, and G. Wang, "On interpretability of artificial neural networks: A survey," *IEEE Trans. Radiat. Plasma Med. Sci.*, vol. 5, no. 6, pp. 741–760, 2021.

[125] W. He, C. Peng, M. Wang, X. Ding, M. Fan, and

H. Ding, "Sensitive attribute recognition and classification algorithm for structure dataset," *Application Research of Computers*, vol. 37, no. 10, pp. 3077–3082, 2020.

[126] R. Latif, S. H. Afzaal, and S. Latif, "A novel cloud management framework for trust establishment and evaluation in a federated cloud environment," *J. Supercomput.*, pp. 1–24, 2021.

Linbin Liu received the B.S. degree in information security from the North China Electric Power University (Baoding), Baoding, China, in 2019. He is currently pursuing the Ph.D. degree with the School of Cyber Science and Engineering, Wuhan University, Wuhan, China. His research interests include privacy protection and industrial big data analytics.

June Li received the B.S. and M.S. degrees in electrical engineering and computer engineering from the Wuhan University of Hydraulic and Electric Engineering, Wuhan, China, in 1986 and 1989, respectively, and the Ph.D. degree in computer engineering from Wuhan University, Wuhan, China, in 2004. She is currently a Professor with the Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University. Her research interests include network architecture, cyber security, blockchain, cyber-physical systems, and security of power industrial control systems.

Jianming Lv (Member, IEEE)received the B.S. degree in computer science from Sun Yat-sen University, Guangzhou, China, in 2002, and the Ph.D. degree from the Institute of Computing Technology, University of Chinese Academy of Sciences, Beijing, China, in 2008. He is currently a Professor with the School of Computer Science & Engineering, South China University of Technology. His research interests include data mining, computer vision, and distributed computing and privacy.

Juan Wang (Member, IEEE) received the B.S. and M.S. degrees in computer science and the Ph.D. degree in information security from the school of Computer Science, Wuhan University, Wuhan, China, in 1998, 2004, and 2008, respectively. She is currently a Professor with the School of Cyber Science and Engineering, Wuhan University. Her research interests include cloud security, trust computing, and SDN and NFV security.

Siyu Zhao received the B.S. degrees in information security from the school of Cyber Science and Engineering, Wuhan University, Wuhan, China, in 2020. His research interests include blockchain and ICS security.

Qiuyu Lu received the M.S. degree from the School of Cyber Science and Engineering, Wuhan University, Wuhan, China, in 2020. She is currently pursuing the Ph.D. degree in the School of Cyber Science and Engineering, Wuhan University. Her research interests include QoS guarantee for communications of smart grid, Industrial Internet, and cyber security.

TABLE IV
SUMMARY OF THE KEY TECHNOLOGIES.

Challenge	Key tech	Solution	Advantage	Disadvantage	Objective	Applicability
How the data intelligent analytics model fit into the new characteristics of IBD?	IBD Fusion and Analytics	Joint Representation	Preserve the hierarchical clustering structure of heterogeneous graphs. Capture local and global features in the graph.	Rely on the integrity of heterogeneous graphs.	Reveal the multi-scale association features of IBD.	Suitable for heterogeneous multi-modal data.
		Fusion Analytics	Utilize the diversity between different modes. Asynchronous curriculum learning strategies can improve efficiency and stability. Low requirements for the quantity and quality of learning materials.	Need appropriate fusion schemes with subjective and empirical factors.	Enhance the domain adaptation and interpretability.	Suitable for scenarios with data distribution differences between source and target domains.
		Distributed Computing	Solve the performance bottleneck of centralized scheduling system. Realize rapid and flexible expansion.	Complex node coordination, consistency, load balancing, etc.	Adapt to distributed and large-scale IBDA scenarios.	Suitable for diverse and complex IBDA scenarios in the distributed Industrial Internet.
How to analyze and utilize data while ensuring privacy and security?	IBD Privacy and Security Protection	Attribute Identification	Identify potential sensitive attributes.	Unable to handle datasets with unknown or inaccurate probability distributions.	Hierarchical IBD privacy protection.	Suitable for sensitive attribute identification for structured datasets.
		Secure FL	Distributed modeling without sharing data. Eliminate dependence on central servers. Reduce single point of failure and trust risks. High security and auditability.	Issues such as system heterogeneity, data heterogeneity, number of participants, malicious participants, etc.	Privacy-preserving IBDA.	Suitable for distributed scenarios where model training requires private data.
		Access Control	Fine-grained and flexible access control. Continuous and dynamic trust evaluation of subjects, objects, and environments.	Complex access control strategies and management strategies.	Secure data access during DS&T.	Suitable for data access control in complex Industrial Internet scenarios.
How to share and trade data efficiently in the Industrial Internet environment without mutual trust?	Blockchain Supporting Secure and Efficient IBDA	Scalable Block	Achieve efficient heterogeneous data retrieval. Reduce storage and transmission costs. High scalability.	Complexity of blockchain. High computation and management overhead.	Store IBD uniformly in the blockchain.	Suitable for scenarios where heterogeneous IBD and models are stored and retrieved.
		Consensus Algorithms	High throughput and efficiency. High reliability and supervisability.	Compatibility and coordination issues between different consensus mechanisms.	Block consensus in different application scenarios in IBDA platform.	Suitable for scenarios with high throughput, efficiency, reliability, and supervisability requirements.
		Smart Contracts	Improve the security and efficiency of FL and DS&T.	Difficulty in development and maintenance.	Ensure the efficient and secure operation of IBD applications.	Suitable for FL and DS&T, and can be extended to other application scenarios.