EasyChair Preprint

№ 8965

Comparative Study on Data Sovereignty Guarantee Technology

Yaodong Tao and Shuai Yang

October 3, 2022

Comparative Study on Data Sovereignty Guarantee Technology

Tao Yaodong
Beijing Jiaotong University, Industrial Internet Security Research Center / Director
Beijing,China
E-mail:taoyd@bjtu.edu.cn

Yang Shuai
Shenyang Ligong University, School of information science and Engineering
Shenyang,China
E-mail:yangshuai@dualpi.com

***Abstract*—As an economic commodity, data sharing, circulation and trading can not only reduce the maintenance and management costs of enterprises, but also tap the potential value of data, improve the internal workflow of enterprises and the cooperation between enterprises. The marketization of data elements and the clarification of data sovereignty are the difficulties that hinder the flow of data at present. This article aims at one of the current data circulation problems: how to maintain data sovereignty, and makes exploration and research in combination with the current era background. For the current research projects and products, compare and analyze the technologies used to maintain data sovereignty. Finally, on the basis of the current technology, it gives suggestions for the development of data sovereignty protection technology in the future.**

Keywords—data sovereignty, data security, trusted circulation of data, usage control

I. INTRODUCTION

With the deepening of digitalization in various industries and the increasing amount of digital storage, how to safely store, effectively use and conveniently manage data has become a difficult problem for most companies on the way to digitalization.

Because of the special nature of data itself, such as non competitiveness and exclusivity, data seems to have become a new economic commodity[i]. This has also created a new ecosystem - digital ecosystem. Like the ecosystem on which human beings depend, digital ecosystem needs the active participation of participants, and data circulation is the key activity for the prosperity of digital ecosystem.

At present, China has set up many Data Exchanges or Data Centers. The purpose of the establishment is to promote the circulation of data, but the implementation is not so success. According to research, the current data transactions have the following problems: difficult to confirm rights, difficult to price, lack of protection technology, difficult to comply with rules and supervision, and insufficient relevant protection laws[ii]. As a result, the willingness of data owners to participate is low, which leads to difficulties in data circulation, and the digital ecosystem is facing a crisis.

Identify applicable funding agency here. If none, delete this text box.

In view of the above problems, some projects have begun to study. For data sovereignty, it is defined in the European project described in Section Ⅱ A: Data sovereignty refers to finding a balance between the need to protect personal data and the need to share data with others.

So about the current research status of this balance, the article launched an investigation. The remaining structure of this article is as follows: Section Ⅱ, briefly introduces the current research projects related to data trading / sharing at home and abroad. Section Ⅲ, according to the research activities, expounds the current needs of safeguarding data sovereignty technology. Section Ⅳ, introduces the relevant products of the domestic enterprises investigated and compares and summarizes the technologies used. Section Ⅴ, summarizes and prospects for the future.

II. RELATED RESEARCH PROJECTS

Up to now, the mainstream of research projects on promoting data sharing / trading and effectively safeguarding data sovereignty are all in Europe. Projects of this kind in Japan and China are starting.

A. European project - International Data Space

It evolved from the Industrial Data Space[iii] project of Fraunhofer Laboratory in Germany to the current International Data Space project (hereinafter referred to as IDS). Its coverage extends from industry to the whole industry, from Germany to Europe.

IDS released its first white paper in 2017. The project is positioned as an open source project, which aims to lay the foundation for the future innovative data economy in Europe. In the same year, the first version of the reference architecture model white paper was published. Its architecture model is shown in Fig 1.

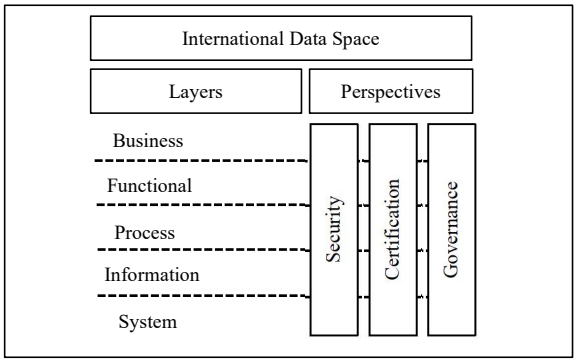

Fig. 1. IDS Reference Architecture Model (Image Source GitHub IDS Architecture 4.0)

The architecture runs through five layers from the perspectives of security, authentication and management, and aims to build a secure, trusted and decentralized data sharing / trading space. The starting point is to maintain the data sovereignty of data owners, promote the trusted circulation of data, and make data a new commodity to realize free sharing / trading.

Fig 2 depicts IDS participants and their role interactions.

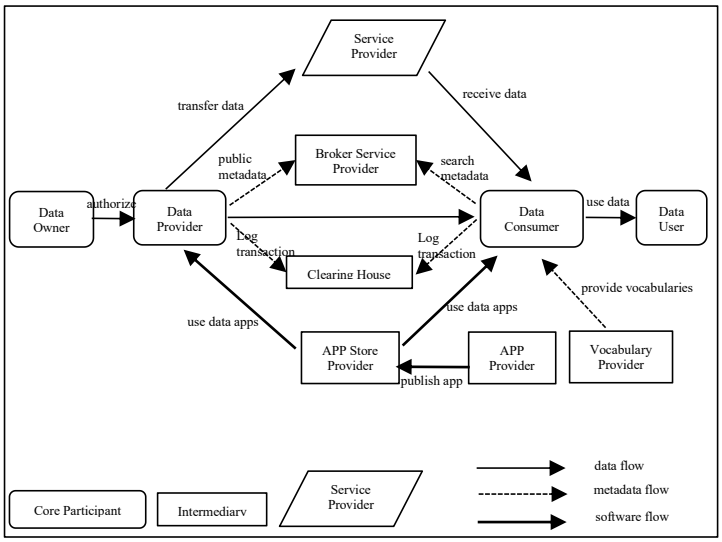

Fig. 2. IDS Participants and Role Interactions(Image Source GitHub IDS Architecture 4.0)

Credibility is the premise of realizing the content drawn in Fig 2. IDS deploys Evaluation and Certification Center. The former is to comprehensively evaluate all organizations that want to participate in the space and send the evaluation results to the latter. The latter issues the X.509 certificate based on this result. The difference lies in the combination of Dynamic Attribute Provisiong Service (DAPS, managing the dynamic attributes of participants) and Dynamic Trust Monitoring (DTM, monitoring network security behavior).

As for the credibility of equipment and environment, IDS is implemented by relying on the core device connector[iv]. Fig 3 is the schematic diagram of connector design.

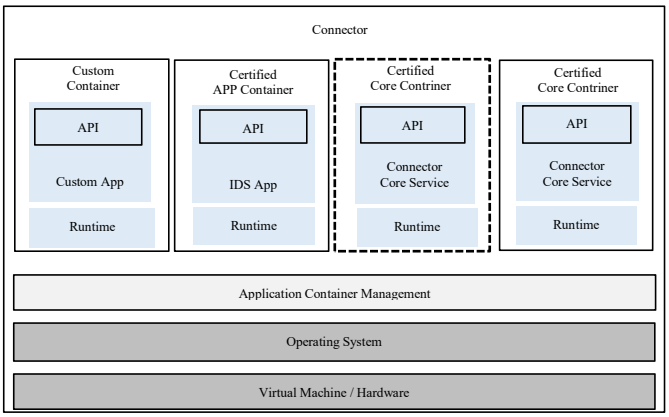

Fig. 3. Connector Design Schematic Diagram(Image Source GitHub IDS Architecture 4.0)

The design of connector adopts container isolation technology to prevent data from falling into the disk of consumer and prevent data leakage. Data provider and consumer can upload or use data in Swaager through only open APIs. The secure transmission of data is a secure channel built through TLS after both sides of the connector confirm their identity certificates.

Another key issue is the safety and controllability of data. In order to solve this kind of problem, IDS adopts the useage control technology which is not as mature as the access control technology.

Compared with access control technology, which has various mature models, such as role-based access control[v], attribute based access control[vi], mandatory access control[vii], etc., useage control technology is still in the research stage.

In 2003, jaehong Park and Ravi Sandhu first proposed UCON (usage control) model at the ACM meeting at that time[viii]. At that time, the amount of network data was not as large as today, and access control, trust management and digital watermarking were enough to meet the security needs of users, so the concept was not widely concerned by the industry at first. There are only a few small areas of discussion.

In 2004, UCON model was further improved into UCONABC model[ix]. It can be seen from Fig 4 that the latter is further improved than the former.

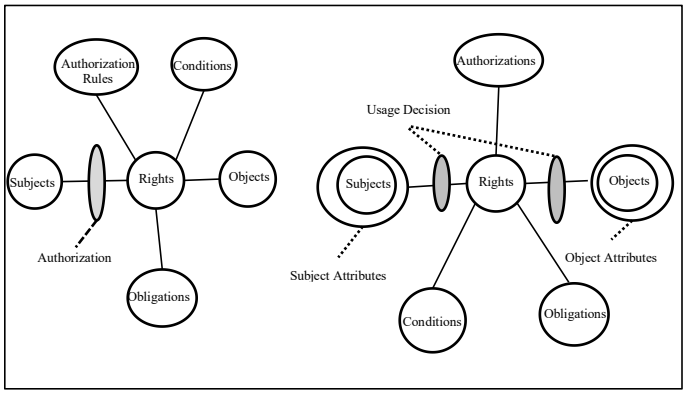

Fig. 4. UCON Model(left) and UCONABC Model(right)

In 2010, Aliaksandr Lazouski of the University of Pisa published a review on useage control[x], which analyzed, compared and summarized the useage control technology in terms of computer security.

Since 2010, the literature on the use of usage control technology is based on UCONABC model, combined with emerging technologies(like cloud computing[xi]) to study .

Before sending data, provider can extend the controllable time of data and ensure data sovereignty by adding control strategies to the data. The use of control strategies will be introduced in Section ⅣA.

B. Japanese project - Connected Industries Open Framework

Japan's Industrial Value Chain Initiative (IVI) is a member of IDS. The organization released the Connected Industries Open Framework(CIOF) in 2019. Its positioning is a collaborative platform that provides or uses valuable data across business institutions and company boundaries. Its purpose is to promote the interconnection between different enterprises and equipments, and realize the safe and reliable circulation of different data.

See Table Ⅰ for the configuration of CIOF system[xii]. It has the following hierarchical system architecture, so that multiple operators can provide and use data on the Internet based on contracts.

TABLE I. CIOF SYSTEM CONFIGURATION

System Configuration	***Function analysis***	***Manage / Provide***
Centralized Server	Authenticate participants	IVI Management
Linked Server	Management platform business and transaction contract	IVI Management
Linkage Manager	Create trading contracts and manage data transactions. Edit Dictionary (dictionary is used to solve the operability problems caused by different formats of data provided by different organizations)	IVI Management
Link Terminal	Communicate with the linked server, similar to connector	IVI Provide
Edge Controller	Directly access the connected link terminal and mediate the data	Provided by CIOF Partners
Service Facilities	Software that plays a role in the business	IVI Provision / Participant R & D

From Table Ⅰ, we can also see some corresponding roles between CIOF and IDS, but at present, CIOF is still in development, and more information has not been announced.

C. Chinese project - Trusted Industrial Data Matrix

At the end of 2021, the China Academy of Information and Communications released the white paper "Trusted Industrial Data Matrix 1.0" marking the beginning of a new era of data circulation and trading in China.

According to the white paper, China is a large manufacturing country, but China's industrial data sharing and circulation industry is still in its initial stage. The non circulation of data has caused great resistance to the upstream and downstream development of enterprises, which not only affects the economy, but also affects resources and other aspects. Therefore, it is necessary to learn from the experience of developed countries and build a framework platform in line with China's national conditions[xiii].

Fig 5 shows the trusted industrial data space architecture. It is divided into five layers, and the design concept is similar to Tcp/Ip protocol. From bottom to top, there are data access layer, transmission processing layer, intermediate service layer, data control layer and data application layer.

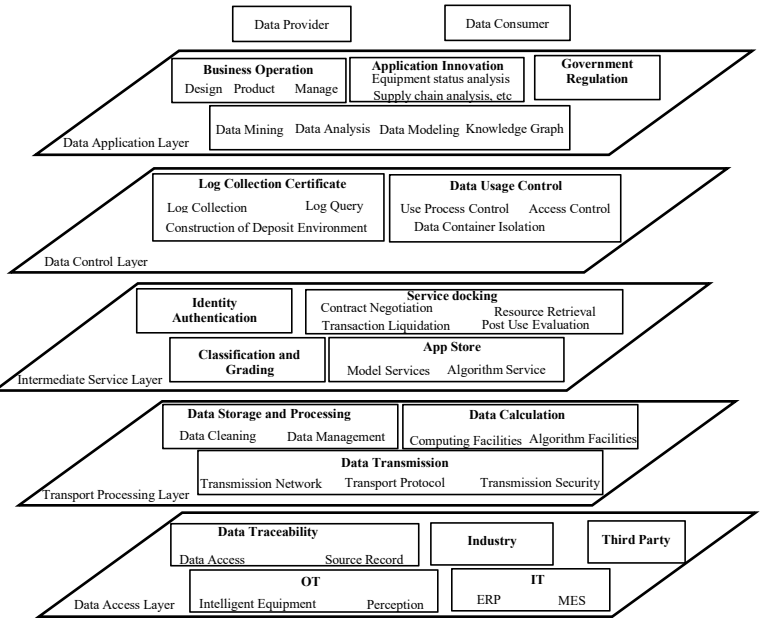

Fig. 5. Trusted Industrial Data Matrix Architecture(Image Source Trusted Industrial Data Matrix 1.0)

In addition, it can be seen from Fig 5 that the data sovereignty assurance technology in Trusted Industrial Data Matrix is reflected in the data control layer. The use of data is monitored in real time through log collection and certificate storage, together with control technology, which realizes the function similar to IDS usage control strategies.

Due to its late start, the Trusted Industrial Data Matrix is still developing. With the participation of more scientific research institutions, colleges and universities, and powerful enterprises, it will certainly attract much attention in the future.

III. TECHNICAL REQUIREMENTS FOR DATA SOVEREIGNTY GUARANTEE

According to the field research and the reading of relevant reports, it is found that at present, the sharing and circulation of production information within the enterprise or between upstream and downstream still adopts the most traditional way - a special person holding a special USB flash disk to transmit information. Undoubtedly, this is the most reliable way

summarized before, but it is also the way that enterprises want to change but dare not change.

After the statistics of the concerns of enterprises in data sharing / trading, the following results can be obtained, as shown in Fig 6.

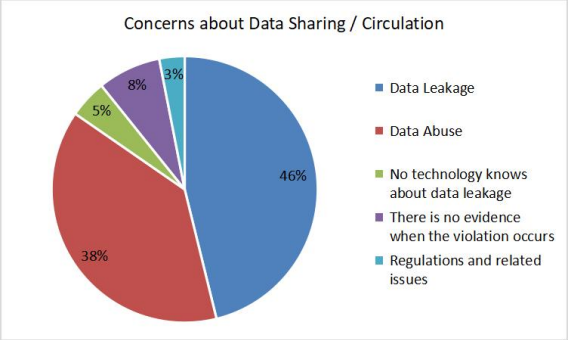

Fig. 6. Statistics on Concerns in Data Sharing / Trading

Among them, data leakage and data abuse are the biggest concerns of enterprises. Solving such problems can greatly promote the circulation of data and speed up the pace of industrial digitalization.

IV. Analysis and Comparison of Domestic Related Production

This section is based on the investigation of relevant data circulation products of domestic enterprises, classifies the products according to the key technologies used, and makes a comprehensive comparison, so as to clarify the advantages and disadvantages of the current domestic technical solutions for maintaining data sovereignty. TABLE Ⅱ shows the comparison results.

TABLE II. Table Type Styles

Cateory	*Centralization*	*Maturity*	*Implementation difficulty*	*Manageability*	*Enterprise*
Control Technology	×	Embryo	Difficult	—	A
Privacy Computing Technology	√	Mature	Medium	Easy	B
	√	Mature	Easy	Easy	C
Process Control Technology	√	Mature	Easy	Easy	D

Cateory	*Domain*	*Operation*	*Operator Requirements*	*Reliability*	*Adopted Technology*
Control Technology	Industry and Finance Medical treatment, etc	API	Professional knowledge required	—	Access Control Usage Control
Privacy Computing Technology	Government affairs and Medical treatment Finance, etc	API	Professional knowledge required	Medium	Federal Learning Secure Multi-Party Computation Ciphertext Calculation Secure Computing Sandbox
	Retail, Government Finance, Medical treatment, etc	User Interface	No professional knowledge required	Medium	Federal learning Secure Multi-Party Computation Trusted Execution Environment Differential Privacy
Process Control Technology	Finance, Energy Medical Treatment , Sales Education , Internet, etc	User Interface	No professional knowledge required	Medium	Trusted Process Process Control

A. Control Technology

The product of enterprise A is still in the development stage, so some relevant options in the table are unknown.

Its design purpose is to build a data space for upstream and downstream enterprises to promote the development of the industry. Similar to IDS architecture, it adopts decentralized architecture. What is special is that the information of participants, the execution of trading contracts and other links are not only signed online, but also reviewed offline.

In order to ensure data sovereignty, usage control technology is adopted. This technology development has been mentioned in Section Ⅱ A, and will not be described here. This technology is still in the development stage, and the vision is beautiful, but there is no unified standard architecture for everyone to use. There are also problems and technical difficulties in the preparation of control strategy language.

As for the usage control strategy languages, at present, IDS has officially announced 21 kinds[xiv]. According to the function of maintaining data sovereignty that can be achieved by the research products, it is matched with IDS 21 kinds of useage control strategies. The matching results are shown in Table Ⅲ.

TABLE III. TABLE TYPE STYLES

Implemented data sovereignty strategy (Mapping with ids 21 use control strategies)	Enterprise			
	A	*B*	*C*	*D*
1 Allow / Prohibit data use	√	√	√	√
2 Data usage is limited to a group of systems or applications		√		
3 Restrict data usage to specific connectors			√	
4 Data usage is limited to a group of users				
5 Restrict the use of data to specific locations				
6 Restrict the use of data for specific purposes				
7 Restrict the use of data in case of specific events				
8 Data usage is limited to the security level of the connector				
9 Limit data usage to specific time intervals	√			
10 Limit the use of data to a specific time range	√			√
11 Do not use these data more than n times	√			√
12 Use data and delete later		√	√	
13 Restrict data usage to specific states				
14 Modify data (transmission)	√			
15 Modify data (static)				
16 Record data usage information	√	√	√	√
17 Notify one party or specific user groups when using data				√
18 Attach policies when distributing data to third parties				
19 Distribute data only when data is encrypted		√	√	√
20 Permanent data sales restrictions				
21 Lease data restrictions				

At present, the research literature on the usage control strategy language is basically XML language[xv]. In this regard, IDS developed Lucon, a machine-readable control strategy programming language similar to natural language. It can be written and generated in Exclipse Lucon file to realize the function of self writing and using control language strategy[xvi]. But there are many problems in practice.

In addition, how to really realize the deletion of data user according to the contract after using data is also a difficult problem that prevents data owner from sharing / trading data in practice. Many organizations hold this view and believe that rather than having no way to know whether the data is deleted, it is better to adopt popular Federal Learning technology to ensure data sovereignty. However, most of the current Federal Learning technology products are used in the financial industry, which may be very uncomfortable for data from other industries, such as industrial device diagrams.

B. Privacy Computing Technology

Such products technically focus on the combination of privacy computing technologies. In the book "privacy computing"[xvii], the author divides this technology into privacy encryption computing technology and privacy protection computing technology. The former focuses on cryptography, key distribution and protection; The latter focuses on the protection of data privacy and strives to achieve the availability and invisibility of data. According to this classification, we can see that the technologies used by enterprises B and C belong to privacy protection computing technology.

1) Privacy Computing Technology - Enterprise B

This product has been mature. The centralized structure is adopted. The data should be uploaded to the intermediate platform first and kept for backup. This undoubtedly increases the risk of data security. Once there is any security problem in the intermediate platform, the consequences will be very serious.

Sandbox can completely isolate the space between different computing tasks, and every operation in the sandbox has a detailed record[xviii]. The product uses sandbox technology to build a central trusted data computing environment. Participants first conduct model training locally through Federated Learning, and transmit the training results to the central trusted data computing environment. The central trusted data computing environment then calculates the ciphertext of the training results, and finally sends them to data consumers. In order to make the data available and invisible.

In addition, for the operation records of data consumer, by using the openness and transparency of blockchain, the data transaction / sharing process can be trusted, controllable and supervised.

Ciphertext computing and Federated Learning require strong computing power, and communication overhead will also be increased in the transmission process, which will undoubtedly reduce the willingness of enterprises to share and circulate data.

From the above brief operation process principle, we can see that the existing data sharing transaction / circulation privacy computing technology scheme (including enterprise C product in the next section) is not like usage control technology. What consumers can use is only a model after federal learning, so the application of this scheme will have limitations. It can be widely used in advertising, precision delivery and other related fields, but it is not applicable to industrial manufacturing.

2) Privacy Computing Technology - Enterprise C

Unlike enterprise B, which uses secure computing sandbox technology to build a trusted computing environment, enterprise C adopts the trusted execution environment (TEE) technology combining software and hardware.

The trusted execution environment of this product adopts Intel SGX, which is widely used in cloud computing system. Taking Intel SGX as the reliable guarantee of intermediate task distribution and processing center. The product design concept adopts management centralization and calculation decentralization[xix]. The centralization of management can be reasonably deployed, optimize efficiency, and implement the credibility of the identity of participants. Decentralized computing can reduce the frequency of data exposure. Participants can train the data model on their own side through Federated Learning, and only need to upload the training results.

However, according to the research literature published in recent years, Intel SGX trusted execution environment still faces many security problems. Because the security of TEE depends on the specific design of the manufacturer, not on the calculation theory of similar algorithms.

In the process of deploying this product, you need to classify your own data assets first. Enterprises will send professionals to target customers for classification work, and finally the classification results will be put on the cloud. Through a friendly user interface, data owner can easily view, manage, trade and share their own data. However, it is time-consuming and laborious to completely rely on manual data classification, and there will be security risks. By combining Machine Learning, training models and using models, it is a better method to rely on manual audit for some data types that are difficult for machines to distinguish.

C. Process Control Technology

The main application scenarios of such products are: for .Doc type files are only allowed to be used within the company and cooperative companies, and the company headquarters has only relevant setting functions, such as limited opening times of files, recording file usage information, file usage validity, etc. These functions map to some of the 21 usage control strategies of IDS.

In this product, the insensible encryption technology is also used, so that the internal employees of the company do not need to decrypt the file and other complex operations when using the file, which greatly improves the convenience. Insensible encryption technology encrypts and restricts the data flow area when it is created. Even if the data is taken away from the set area by social software or USB flash drive, it cannot be decrypted.

As for how to realize the limited use of data, the product updates the real-time operation data by using blockchain contract technology. In addition, the contract technology is also used to register and store trusted processes, and change the name of trusted processes to hash value, so that only trusted processes can use data.

Although trusted process technology and senseless encryption technology can greatly reduce the risk of data leakage, there is no absolutely safe technical solution after all. This product cannot prevent attackers from obtaining the contents of files through processes, remote code execution, or remote memory reading. In addition, in case of power failure, it may also cause data damage.

V. SUMMARY AND OUTLOOK

Comprehensively investigate the technology used in the product, analyze its limitations and summarize its maturity. See Table Ⅳ technical summary table for the results.

TABLE IV. TABLE TYPE STYLES

Serial Number	Technology	Limitations	Time		
			Within 3 years	***3 to 5 years***	***Five years later***
1	Usage Control	The technology is immature, limited to theoretical research, and there are few landing products.	The technology is further mature, and the number of landing applications is increased	Promotion and Application	Universal application
2	Access Control	Mature technology and weak data control performance cannot meet the current demand for maintaining data sovereignty.	—	—	—
3	Federal Learning	The transmission of intermediate calculation results increases the communication overhead and cannot ensure the authenticity of the data provided by the participants.	Create more examples in areas other than finance	Promotion and Application	Universal application
4	Secure Multi-Party Computation	The computing speed is 6 orders of magnitude slower than plaintext computing, and the communication overhead is large. In addition, it is sensitive to delay.	Increase of landing instances in restricted scenes	Application examples of promotion and development	It is estimated that it will take 10 years to reach the peak of technological maturity
5	Ciphertext Calculation	The requirement of computing power is high, and the corresponding cost is also high. The universality is not strong, and the scope of application is small.	Landing of mature solutions for relevant scenes	Promotion and Application	Universal application
6	Trusted Execution Environment	The safety of products cannot be fully guaranteed.	There have been many instances in different scenarios	Examples generated by combination with blockchain	Promotion and application examples
7	Secure Computing Sandbox	The construction cost is relatively high. And in the face of a large number of data processing, the speed is relatively slow.	There are mature container lightweight schemes	—	—
8	Differential Privacy	According to the calculation results, the noise is increased to reduce the risk of eavesdropping with a certain probability, but at the same time, the calculation accuracy is reduced. In addition, it is not applicable to scenarios that require high transparency and interpretability of privacy protection. Lack of technical personnel.[xx]	It has mature applications in statistics and data query	Examples of application scenarios combined with machine learning algorithms are further increased	It is estimated to reach the peak of technological maturity in 10 years
9	Trusted Process	Mature technology	—	—	—

In enterprise B of VI, we mentioned that the usage control technology is aimed at the control of data itself, which has a wide range of applications, while privacy computing focuses on encryption and decryption and the availability and invisibility of data, which is not applicable to some scenarios such as industrial manufacturing. However, the usage control technology also lacks the function of desensitizing data like privacy computing technology. Once the data is leaked, the consequences of the former are more serious than the latter.

At present, there are not many examples of process control technology in data transaction / sharing, which need to be customized according to customer needs.

According to the analysis of the current technology development trend, the usage control technology is still the most likely solution to fully realize the maintenance of data sovereignty in the future, which does not rely on hardware and is convenient for deployment and implementation. In the future, the research on usage control technology can start with the unification of the standards of usage control strategies, build a standard usage control language and form a standardized language set, which is very important for usage control technology.

REFERENCES

[i] Otto, B.; Lohmann, S.; Steinbuss, S.; Teuscher, A.IDS Reference Architecture Model Version 3.0; Technical Report; Fraunhofer: Munich, Germany, April 2019.

[ii] Sun Ke.Problems and reflections on the development of value of data elements;Information and Communications Technology and Policy;2021,47(6):63-67.

[iii] Otto, B.; Lohmann, S.; Steinbuss, S.; Teuscher, A. IDS Reference Architecture Model; Technical Report; Fraunhofer: Munich, Germany, 2017.

[iv] IDS Reference Architecture Model Version 4.0.Avaliable Online:https://github.com/International-Data-Spaces-Association/IDS-RAM_4_0.

[v] Elisa Bertino.RBAC models — concepts and trends.Computers & Security;Volume 22, Issue 6, September 2003, Pages 511-514.

[vi] E. Yuan;J. Tong.Attributed based access control (ABAC) for Web services.IEEE International Conference on Web Services (ICWS'05).11-15 July 2005.

[vii] Mandatory-Access-Control Model. In: Access Control Systems. Springer, Boston, MA.2006.

[viii] Sandhu, R.; Park, J. Usage Control: A Vision for Next Generation Access Control. In Computer Network Security, Proceedings of the 2nd International Workshop on Mathematical Methods, Models, and Architectures for Computer Network Security, MMM-ACNS 2003, St. Petersburg, Russia, 21–23 September 2003; Springer: Berlin/Heidelberg, Germany, 2003; pp. 17–31.

[ix] J. Park and R. Sandhu. The UCONABC usage control model. ACM Trans. Inf. Syst. Secur., 7(1):128–174, Feb. 2004.

[x] Lazouski, A.; Mancini, G.; Martinelli, F.; Mori, P. Usage control in cloud systems. In Proceedings of the 2012 International Conference for Internet Technology and Secured Transactions, London, UK, 10–12 December 2012; pp. 202–207.

[xi] Lazouski, A.; Mancini, G.; Martinelli, F.; Mori, P. Usage control in cloud systems. In Proceedings of the 2012 International Conference for Internet Technology and Secured Transactions, London, UK, 10–12 December 2012; pp.202–207.

[xii] CIOF System Configuration.Avaliable Online:https://community.ciof-ivi.com/

[xiii] Trusted Industrial Data Matrix 1.0. China Academy of Information and Communications Technology.Avaliable Online:http://www.aii-alliance.org/index/c145.html

[xiv] Steinbuss S. et al. (2021): Usage Control in the International Data Spaces. International Data Spaces Association.Avaliable Online:https://internationaldataspaces.org//wpcontent/uploads/dlm_uploads/IDSA-Position-Paper-Usage-Control-in-the-IDS-V3.pdf.

[xv] INES AKAICHI;SABRINA KIRRANE.Usage Control Specification, Enforcement, and Robustness: A Survey.arXiv:2203.04800v1 [cs.CR] 9 Mar 2022.

[xvi] How to build yourself data usage control policy.Avaliable Online:https://international-data-spaces association.github.io/DataspaceConnector/

[xvii] Chen Kai;Yang Qiang.Privacy-Preserving Computing.Beijing:Publishing House of Electronics Industry,Feb 2022.

[xviii] William Wright；David Schroh.The Sandbox for analysis: concepts and methods.Proceedings of the SIGCHI Conference on Human Factors in Computing SystemsApril 2006 Pages 801–810.

[xix]The documentation of DataTrust Product.Avaliable Online:https://help.aliyun.com/document_detail/208028.html?spm=a215hz.22503840.0.0.279169fdJXXrGy

[xx] Bart Willemsen.Hype Cycle for Privacy, 2021. Published in 13 July 2021.