Trusted Data Space Technology for Data Governance in the IOWN Era

Tomohiro Inoue and Tetsushi Morita

Abstract

We describe the components of the Trusted Data Space, the data-distribution infrastructure that will support society when the Innovative Optical and Wireless Network (IOWN) is widespread and all information is used by digital twins and artificial intelligence. We also describe a group of technologies (data sandbox technology, secure computation technology, secure matching technology) for achieving governance over data processing by conducting calculations in an encrypted state.

Keywords: data governance, secure computation, confidential computing

1. Background

As the Internet of Things and artificial intelligence (AI) technologies advance, the construction of digital twins that reproduce real-world systems in cyberspace and analyze and predict system behavior is progressing. The construction of digital twins for specific applications is leading the way, which will be interconnected, leading to increased data sharing and data analysis across organizational and industry barriers. For example, to enable various use cases expected in smart cities, such as disaster prevention, crime prevention, and attractive town development, data obtained from transportation operators, restaurants, and entertainment facilities need to be used in addition to public data by linking them with one another.

We are engaged in the research and development of the Trusted Data Space with the aim of creating a world in which a wide variety of data generated by various individuals and companies can be effectively used across the boundaries of organizations and industries and in which everyone can share data with one another, analyze them, and create data with a new purpose in a chain fashion to discover the value of one another's data and maximize the data's value for society as a whole.

The Trusted Data Space (**Fig. 1**) consists of various functionalities for interconnecting various data managed and shared by each data space[*1] to enable data distribution across data spaces; virtual-secure-data-lake functionality for virtually integrating data managed by each data owner; functionality to support value creation through data distribution by matching between data owners and data users through data cataloging, credit evaluation, etc.; and functionality for managing agreements on terms of use in addition to data-access rights and limiting methods of data processing on the basis of those terms of use. This article details the last functionality, which restricts how data are processed.

2. Governance technologies for how data are processed

In a society envisioned by the Trusted Data Space, data are not managed by a centralized operator but assumed managed by each company, organization, or individual, who keeps their data under their control. Ideally, this means that data can be placed where the owner wants it (such as in a trusted device or datacenter), shared with others only when and to the extent necessary, and leveraged only under conditions

*1 Data space: A community with systems and mechanisms designed to enable highly secure data sharing. For details, see the article in this issue, "Data Governance for Achieving Data Sharing in the IOWN Era."

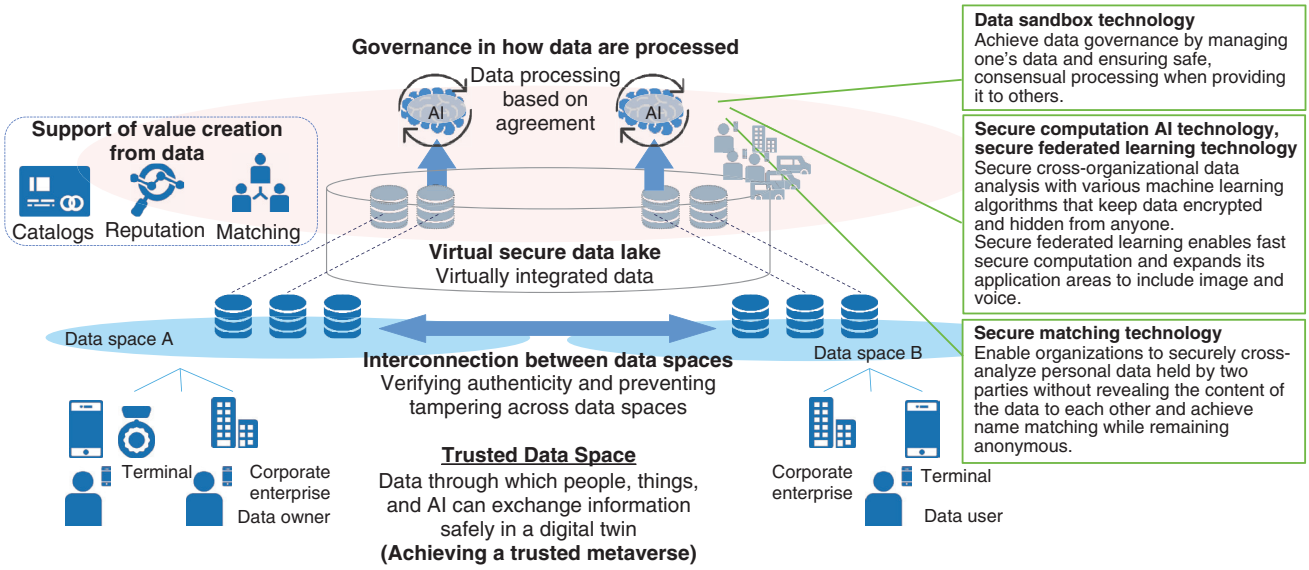

Fig. 1. Trusted Data Space.

prescribed by the parties involved. The virtual-secure-data-lake functionality of the Trusted Data Space enables data that remain under the control of the data owner to be virtually consolidated and retrieved as one huge data lake. Data owners share data-handling policies (e.g., period of use, permission to reproduce, feasible processing), and data users can use the data to the extent permitted by agreeing to the policies. We believe that by establishing such a system, we can create a world in which we can create value in a cascading manner by making it easier to use highly sensitive data that would otherwise be difficult for others to share and by promoting the secondary use of data.

Among the functions related to data governance, technology that uses cryptography to impose certain restrictions on how data are processed has been an area of remarkable development. Encryption technology has been primarily used to prevent theft from third parties when transferring and storing data. With the development of technology that can process data during the computation process while keeping them encrypted, the speed and practicality of computation have increased, making it possible to prevent data during the computation from being stolen and used for purposes other than their original purpose. Analysis operations using personal data and corporate trade secrets can be conducted without leaking data or "looking inside the data." This enables safer data processing, as well as new integrated analysis across companies and industries that share data that were previously difficult to disclose with other organizations.

We describe each of these governance technologies in the following sections.

3. Data sandbox technology

As mentioned above, the cascading creation of value through data utilization requires cross-organizational data collaboration. However, such utilization has not spread widely because in actual businesses, there is significant concern that data once shared by others may be replicated or used for purposes other than what was intended. To address this concern, we developed a technology (data sandbox technology: **Fig. 2**) that enables companies and organizations to hide the knowledge they manage (in this article, this means specific information, such as data or algorithms, that the organization wants to keep secret) from one another while leveraging the combined value. This technology can prevent duplication and abuse of shared knowledge by carrying out processing within a special trusted execution environment (TEE) provided by modern central processing units (CPUs).

With this technology, an isolated processing execution environment consisting of a TEE called a data sandbox (DSB) is created on the platform, and data processing is executed in it. (1) When data owners

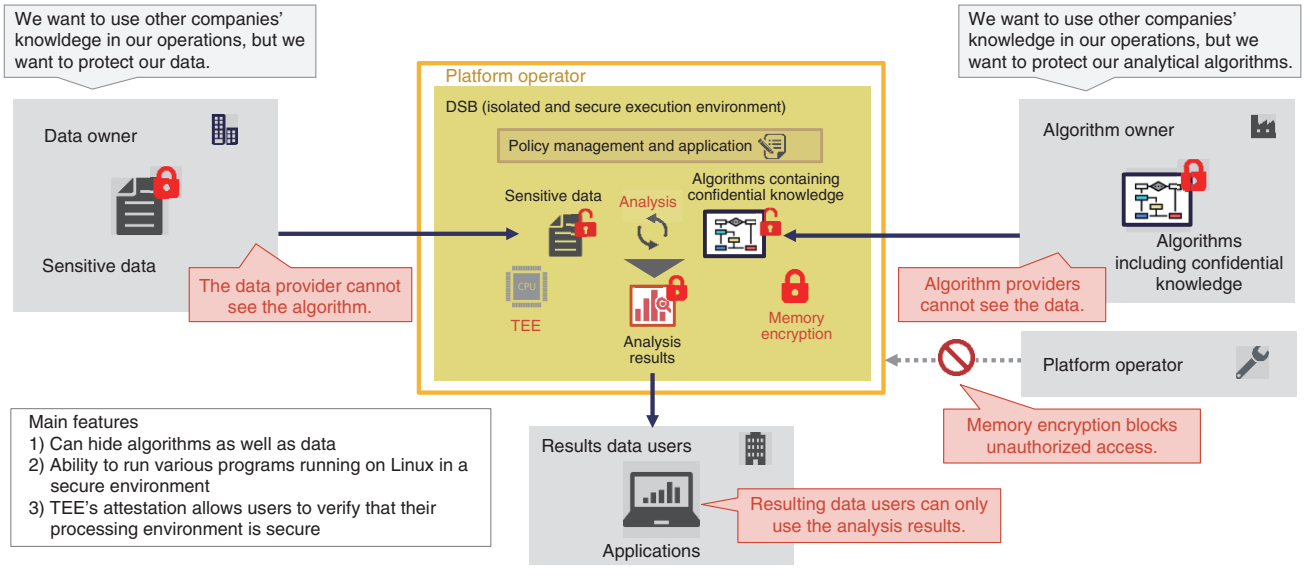

Fig. 2. Data sandbox technology.

and algorithm owners agree to share each other's knowledge and register policies on the platform, DSBs are generated. A DSB is restricted so that it cannot communicate with the outside world, the memory disk is encrypted, and neither the operating system nor the operator can see inside. (2) Each data owner and algorithm owner generates and shares a symmetric encryption key with the DSB and places data and algorithms encrypted with their key in the DSB. Each owner can verify whether the DSB was created in accordance with the policy agreed upon in advance, that is, whether it was not swapped with malicious data or algorithms, by referring to the attestation report[*2] provided by the system. (3) The DSB uses a symmetric key between the data owner and algorithm owner to decrypt the data and the algorithm and execute the operation. (4) The DSB encrypts and returns the processing results with a symmetric key created and shared with the data user. (5) After processing, data and algorithms are deleted along with the DSB. With such a mechanism, the data sandbox technology makes it possible to use data without anyone having access to the processing process and result data as well as input data and algorithms.

4. Secure computation technology

Secure computation is a technology that allows data to be computed while remaining encrypted consistently even within the CPU. In addition to encryption during data communication and storage, secure computation can also be executed during data computation without ever decrypting the data, ensuring a high level of security.

NTT's secure computation uses a secret sharing scheme that conforms to the ISO (International Organization for Standardization) standards as its encryption mechanism and uses multiparty computation techniques based on the secret sharing scheme. Multiparty computation executes secure data processing while the data are encrypted by executing special cryptographic operations and exchanging encrypted data among multiple servers in accordance with preliminarily defined procedures. While performing these steps, the process is executed "without looking inside the data" because the data are always treated as encrypted pieces called a share in the context of a secret sharing scheme. Secure computation technology is currently available from NTT Communications in the form of the commercial secure computation cloud service SeCIHI (Secure Computation and Information Handling Interface). We are engaged in research and development of a technology that will advance secure computation and enable rapid learning and inference of AI.

*2 Attestation report: A function of TEE that certifies the state (e.g., binaries, settings, etc.) in the TEE based on the trustworthiness of the hardware security chip.

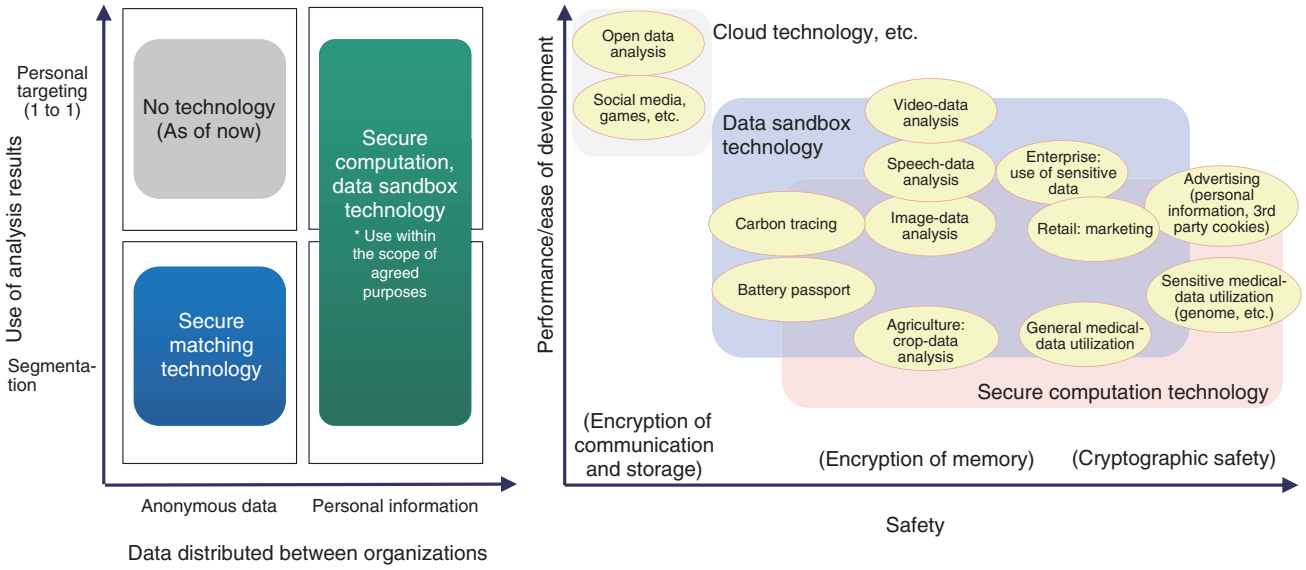

Fig. 3. Application areas of each technology.

5. Secure matching technology

Among the types of data utilization expected in the Trusted Data Space, sharing and analyzing data across industries and organizations for a common target is particularly important and desired. However, from the viewpoint of confidentiality and personal-information protection, data in an organization cannot be freely shared with others for statistical analysis. Secure matching technology enables secure cross-analysis of data by anonymously collating personal information and business data held by an organization between two parties without disclosing the content to the other party. Whereas the data sandbox technology and secure computation technology can execute arbitrary data processing, secure matching technology focuses on data exchange and aggregation processing between two parties, so that each data owner can safely conduct integrated analysis by introducing a simple system into their company and exchanging data.

There are two technological advances in secure matching. The first is a protocol for securely combining and totaling data while keeping them encrypted using advanced cryptography such as commutative hash functions and homomorphic encryption. The other is differential privacy, which adds noise at high speed while maintaining encryption to protect privacy so that the original data will not be known even from the aggregate results. We want to combine secure matching technology with other secure data-distribution technologies developed by NTT to create the Trusted Data Space for reliable data distribution. Secure matching technology is also being used in demonstration experiments at Japan Airlines (JAL), JAL Card, and NTT DOCOMO to improve customer experience and use data to solve social issues.

6. Technology selection for secure data utilization

There is much overlap in the use cases covered with the above technologies. Ultimately, if there were a technology that could theoretically process sensitive data securely and quickly and pass only the necessary results to those who need them, all use cases could be handled. However, no such technology currently exists. Therefore, better technologies need to be determined and applied as appropriate in accordance with the characteristics of the target data, requirements of the parties involved, and legal regulations. **Figure 3** illustrates the different uses of technology that we consider. Secure matching technology can be used when one wants to conduct statistical analysis on a dataset with identifiers (IDs) that can be collated and combined with data from other organizations. This is especially useful when conducting marketing analysis while maintaining the anonymity of personal information. While secure computation is highly secure because it executes operations on the basis of

cryptographic theoretical security, data sandbox technology is characterized by its high processing speed and ability to execute algorithms almost without modification because it uses hardware-based memory encryption. Therefore, a processing method needs to be determined that balances safety and performance as required by the use case.

7. Future developments

To achieve the Trusted Data Space to distribute data in which new value is created in a cascading manner, progress must be made in not only the technologies introduced in this article but also various other areas such as searching and matching reliable data and analysts, standardizing data formats, promoting interconnection with existing data spaces, and creating value through AI. We will continue to promote these research and development efforts through collaboration and technical verification with a wide range of partners.

Tomohiro Inoue
Senior Research Engineer, Supervisor, NTT Social Informatics Laboratories.
He received a B.E. and M.E. from the University of Tokyo in 1999 and 2001. He joined NTT in 2001 and is currently engaged in data hub, trusted data-sharing infrastructure, and global data-space interconnection as the leader of Information Sharing Architecture Group.

Tetsushi Morita
Senior Research Engineer, Supervisor, NTT Social Informatics Laboratories.
He received a B.E. and M.E. from Kyoto University in 1996 and 1998 and Ph.D. in engineering from the University of Tsukuba, Ibaraki, in 2010. From when he joined NTT Software Laboratories in 1998 until 2012, he was engaged in research on information retrieval and personalization systems. His current research interests include secure data-sharing technologies.