Next-generation Data Hub Technology for a Data-centric Society through High-quality High-reliability Data Distribution

Seiichiro Mochida and Takahiko Nagata

Abstract

NTT Software Innovation Center is researching and developing next-generation data hubs as part of the Innovative Optical and Wireless Network (IOWN) proposed by NTT. These data hubs will safely distribute diverse types of data including confidential data through advanced means of data protection in an environment of high-frequency/large-capacity data traffic. This article outlines the problems involved with data distribution and introduces the technologies for configuring data governance, the main function of a data hub.

Keywords: data distribution, data-centric society, data governance

1. NTT's goal of a data-centric society

There has recently been an expansion of Internet of Things (IoT) devices thanks to advances in sensing technology and an increase in data sources having broadband connectivity through the launch of 5G (fifth-generation mobile communication) networks. In addition, advances in artificial intelligence (AI) technologies are enabling high-speed processing of data far exceeding the cognitive and processing abilities of humans. As a result, the amount of data generated throughout the world is increasing steadily in a manner that is expected to not only continue but accelerate in the years to come.

In addition to using data only within closed organizations such as companies, NTT seeks to achieve a data-centric society in which massive amounts of data will be widely distributed beyond the traditional borders of industries and fields at ultrahigh speeds between autonomously operating AI-based systems. This society will enable the creation of totally new value and development of solutions to social problems through novel combinations of data and expertise.

2. Problems in achieving a data-centric society

However, there are two main problems that must be solved to achieve this data-centric society.

2.1 Limits imposed by current data processing architecture

Data processing is currently executed by individual systems (silos) that differ in terms of purpose and processing method. This silo-oriented architecture results in many copies of the same data in those systems. In a data-centric society in which large volumes of data much greater than current levels are exchanged at ultrahigh speeds between many more entities (humans, systems, devices, etc. that distribute data) than today, this situation will only accelerate, which means that the following problems will arise if the current data-processing architecture continues to be used without modification.

- Storage and network performance/capacity will come under pressure.
- The number of data-processing flows that must be managed will increase, making data management

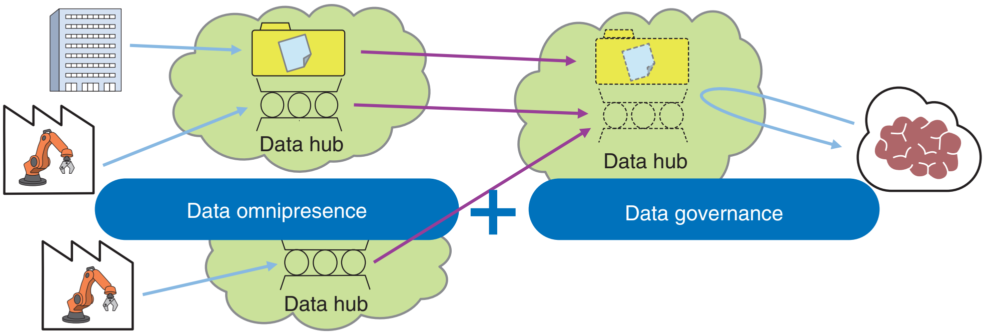


Fig. 1. Next-generation data hubs.

difficult.

- The increase in quantity and types of data items to be processed will make the management of data generation and change history difficult.
- Duplicate and derived data not covered by the rights of source-data providers or by lifecycle management will increase.

2.2 Resistance to sharing confidential information and expertise with other companies

When sharing confidential information and expertise beyond the borders of an organization such as a company, it is common to restrict secondary distribution of shared content or use of that information for purposes other than that agreed upon on the basis of a non-disclosure agreement. However, there are no effective technical mechanisms for preventing secondary distribution or unintended use, so there are limits to enforcing compliance with such an agreement. This has the possibility of hindering a surge in data distribution beyond the traditional borders of industries and fields.

3. Initiatives toward problem solutions

NTT Software Innovation Center aims to solve these problems and achieve a data-centric society by collaborating with various research laboratories including NTT Secure Platform Laboratories to develop next-generation data hubs having the following features (**Fig. 1**).

3.1 Data omnipresence

- Once a data provider creates a folder or queue within a data hub and places data in such a location, authorized users can then use those data from anywhere in the world.
- The data user can perform various types of workload processing or long-term storage with respect to data on the data hub at reasonable cost without having to worry about moving the data around.
- The data user can process data on a data hub by using diverse computing resources having a network connection to that data hub without having to worry about moving or duplicating data.
- The data user can use the application programming interfaces (APIs) of products and services of major storage systems and message broker systems to access a data hub.

3.2 Data governance

- The data provider can maintain and continue to exercise its management rights (governance) in access control, data deletion, etc. with respect to data submitted to the data hub and its duplicated and derived data.
- The data provider can prevent the use of data for purposes other than that agreed upon.
- The data provider can verify who used the provided data for what purpose and for how much.

Between these two major functions of data omnipresence and data governance to be provided by next-generation data hubs now under development, the following section mainly describes the technologies for configuring data governance, which are being developed first.

4. Technologies for configuring data governance

Among the main technologies for configuring data governance, this section mainly describes the data sandbox technology under joint development by NTT Software Innovation Center and NTT Secure Platform

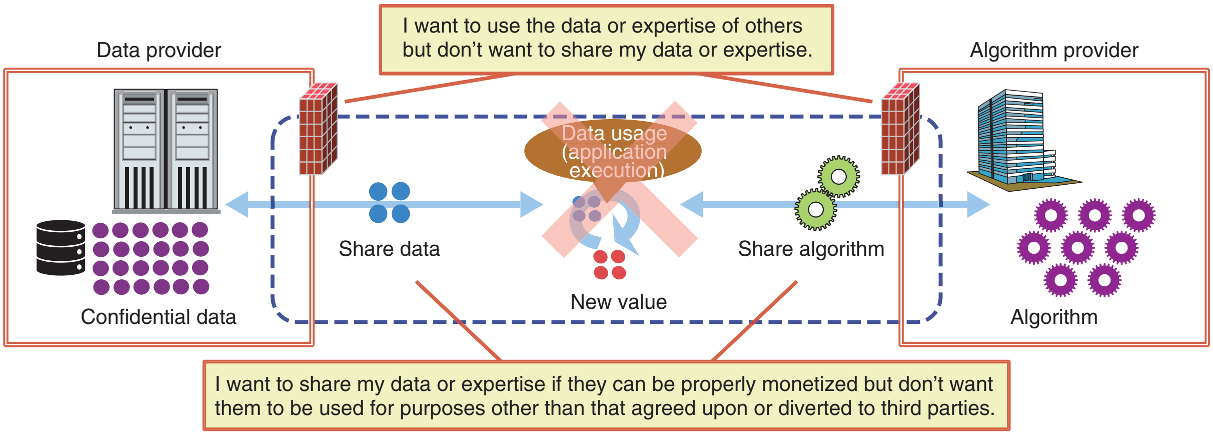

Data provider
I want to use the data or expertise of others but don't want to share my data or expertise.
Algorithm provider
Data usage (application execution)
Share data
Share algorithm
Confidential data
New value
Algorithm
I want to share my data or expertise if they can be properly monetized but don't want them to be used for purposes other than that agreed upon or diverted to third parties.

Fig. 2. Problems with data distribution between companies.

Laboratories. This is followed by brief descriptions of mutual-authentication/key-exchange technology and secure-computation AI technology now under development by NTT Secure Platform Laboratories.

4.1 Data sandbox technology

On achieving a data-centric society in which data and expertise (i.e., for creating an algorithm for giving data value) circulate beyond the borders of an organization such as a company, and new value is obtained by combining those data and expertise, the following can be envisioned in the minds of parties that distribute data and expertise.

- I want to use the data or expertise of others but don't want to share my data or expertise.
- I want to share my data or expertise if they can be properly monetized but don't want them to be used for purposes other than that agreed upon (what type and range of data to be processed by what type of algorithm) or be diverted to third parties.

The only way for individual companies to address concerns like these was to draw up and conclude a non-disclosure agreement—a time-consuming process—and place trust in each other. This type of countermeasure, however, has the potential of hindering the distribution of data beyond the organization (**Fig. 2**).

The question, then, is how to provide a systematic and technical means of defense against the above concerns instead of a defense based on agreements and mutual trust. The answer is the data sandbox technology developed by NTT Software Innovation Center.

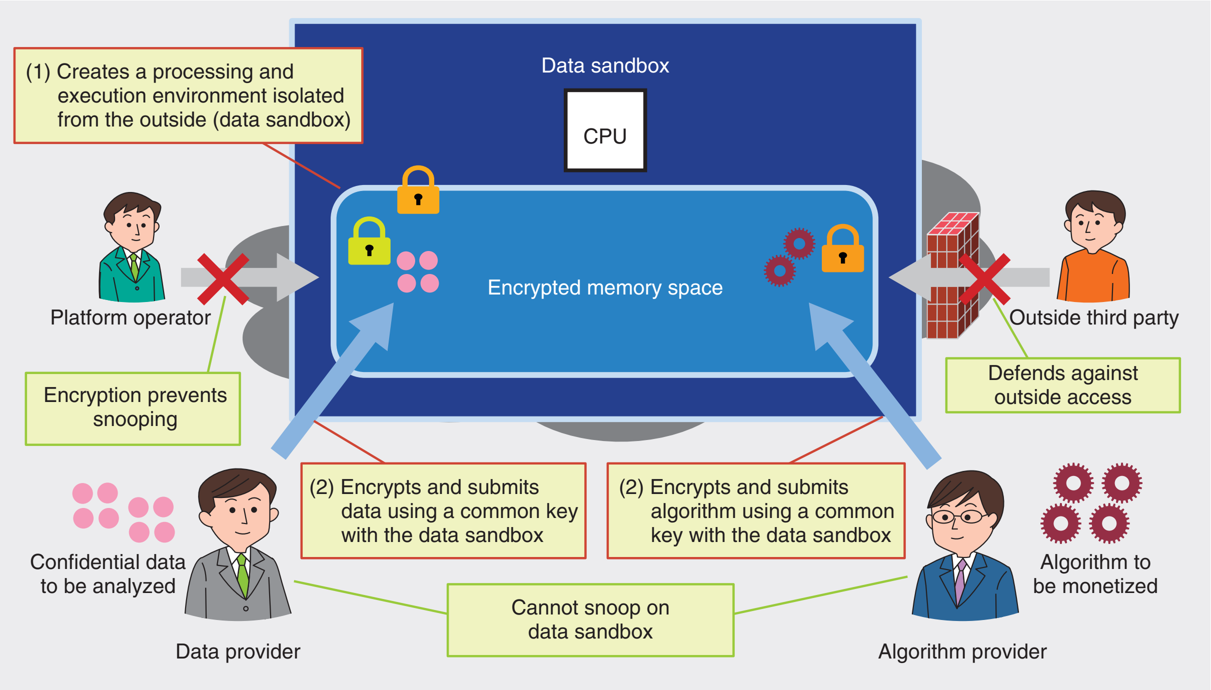
The operation of data sandbox technology can be summarized as follows.

(1) An isolated processing and execution environment called a data sandbox is created on a third-party platform such as a cloud operator that is not a party to data distribution. The data sandbox appropriately restricts communication with the outside and encrypts memory/disk space. The platform operator cannot break this encryption (**Fig. 3**).

(2) Given a data provider having data to be analyzed using another company's algorithm and an algorithm provider monetizing algorithms and providing them to another company, each of these parties generates a common key with the data sandbox and places the encrypted data or algorithm in the data sandbox using that common key. The common key used by the data provider differs from that used by the algorithm provider, thereby preventing the viewing of each other's data or algorithm. The data sandbox also restricts communication with the outside to prevent the data provider and algorithm provider from looking inside the data sandbox while allowing each to only input its data or algorithm (Fig. 3).

(3) The data sandbox decrypts the data and algorithm using common keys with the data provider and algorithm provider. Memory/disk space in the data sandbox is encrypted, which prevents the platform operator from viewing the data or algorithm (**Fig. 4**).

(4) The data sandbox performs processing using the data and algorithm. At this time, the data


Data sandbox
CPU
Encrypted memory space
(1) Creates a processing and execution environment isolated from the outside (data sandbox)
Platform operator
Encryption prevents snooping
Outside third party
Defends against outside access
(2) Encrypts and submits data using a common key with the data sandbox
(2) Encrypts and submits algorithm using a common key with the data sandbox
Confidential data to be analyzed
Data provider
Cannot snoop on data sandbox
Algorithm to be monetized
Algorithm provider

CPU: central processing unit

Fig. 3. Data sandbox (operation overview 1).

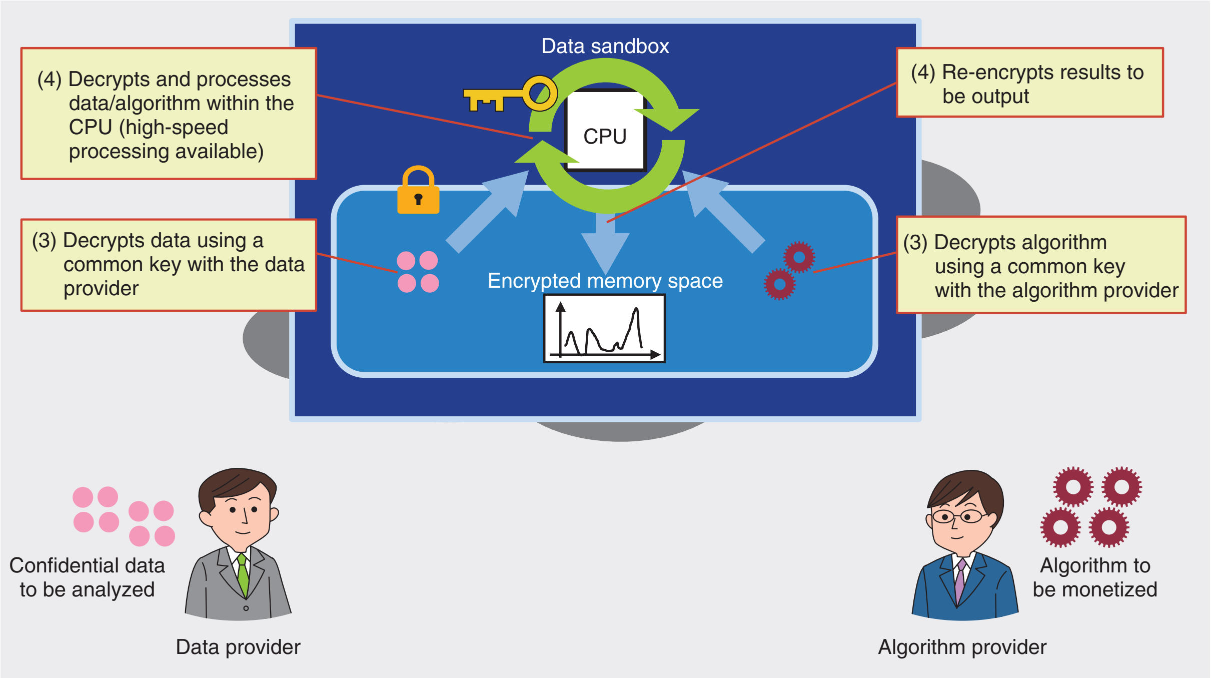

Data sandbox
CPU
Encrypted memory space
(4) Decrypts and processes data/algorithm within the CPU (high-speed processing available)
(4) Re-encrypts results to be output
(3) Decrypts data using a common key with the data provider
(3) Decrypts algorithm using a common key with the algorithm provider
Confidential data to be analyzed
Data provider
Algorithm to be monetized
Algorithm provider

Fig. 4. Data sandbox (operation overview 2).

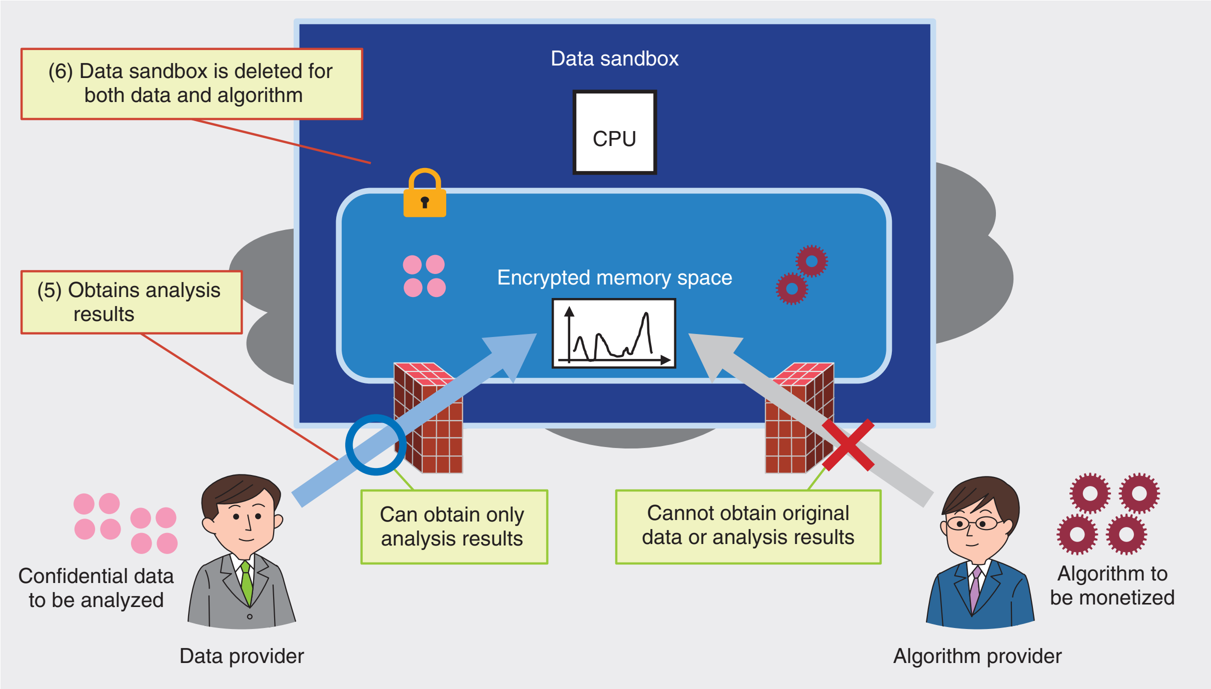

(6) Data sandbox is deleted for both data and algorithm
Data sandbox
CPU
Encrypted memory space
(5) Obtains analysis results
Can obtain only analysis results
Cannot obtain original data or analysis results
Confidential data to be analyzed
Data provider
Algorithm to be monetized
Algorithm provider

Fig. 5. Data sandbox (operation overview 3).

and algorithm in memory/disk space are decrypted only within the central processing unit (CPU) and processed in plain text to enable high-speed processing. The results of processing are again encrypted in memory/disk space when leaving the CPU (Fig. 4).

(5) The data sandbox returns only the processing results to the data provider. Since communications with the outside are appropriately restricted, the algorithm provider cannot get hold of the original data or analysis results even if it willfully or mistakenly submits a malicious algorithm (**Fig. 5**).

(6) The data sandbox is deleted with the data and algorithm after completing processing (Fig. 5).

A party with access to data distribution that uses a data sandbox operating as described above can benefit in the following ways:

- It can use data or expertise of other parties and obtain new value without having to share its data or expertise.
- It can monetize its data or expertise by systematically and technically defending against use for purposes other than that previously agreed upon or diversion to third parties.

4.2 Mutual-authentication/key-exchange technology

Next-generation data hubs will make it possible to instantly share data located anywhere in the world, but they will also require the means of encryption so that only data users approved by the data provider will be able to reference that data. Mutual-authentication/key-exchange technology enables a data provider and data user to verify each other's identity, attributes, etc. and exchange keys for encrypting and decrypting data without being observed by other parties. This enables safe data sharing only with desired parties.

We can expect next-generation data hubs to be connected to IoT devices that will exist in vastly greater numbers than today, which will make it possible to provide massive amounts of real-world data. For this reason, we are developing mutual-authentication/key-exchange technology that requires a minimal amount of computing resources and transmission bandwidth. We can also expect many parties connected to next-generation hubs to provide and use data in a mutually interactive manner. To handle this scenario, we are developing this technology to efficiently execute mutual authentication and key exchange not on a one-to-one basis but among many parties and that can flexibly update keys in accordance with the increase or decrease in the number of parties sharing and using data.

4.3 Secure-computation AI technology

Even if a safe execution environment can be assumed, there will still be not a small amount of data for which decryption is not allowed due to data providers that are uneasy about data decryption or legal restrictions. Secure-computation AI technology enables training and prediction through machine learning with absolutely no decryption of encrypted data. It enables execution of the entire flow from data registration and storage to training and prediction without disclosing the content of that data to anyone. The end result is safe distribution and use of corporate confidential information or information restricted due to privacy concerns. This technology also makes it possible to combine and use data from multiple providers or different types of data in encrypted form. We therefore expect secure-computation AI technology to not only improve the safety of original data but to also enable new value to be uncovered due to an increase in the types and quantities of data targeted for analysis.

5. Toward the future

This article focused on the data-governance function of next-generation data hubs we are now developing. Going forward, we are looking to accelerate this development initiative toward the practical deployment of next-generation data hubs through the development of data omnipresence, which is the other major function of next-generation data hubs, seamless linking of data omnipresence and data governance, and achievement of large-capacity and low-latency capabilities by linking with the All-Photonics Network, a major component of IOWN (Innovative Optical and Wireless Network).

Seiichiro Mochida
Senior Research Engineer, IoT Framework SE Project, NTT Software Innovation Center.

He received a B.E. in science and engineering from Waseda University, Tokyo, in 2002 and an M.E. in engineering science from the University of Tokyo in 2004. He joined NTT Information Sharing Platform Laboratories in 2004, where he was involved in R&D of high-reliability systems. He moved to NTT Communications in 2008, where he worked on 050 IP phone services. He moved to NTT Software Innovation Center in 2012. His current research interests are focused on the system architecture of the IoT management platform.

Takahiko Nagata
Senior Research Engineer, IoT Framework SE Project, NTT Software Innovation Center.

He received a B.E. and M.E. from the Faculty of Instrumentation Engineering, Hiroshima University in 1993 and 1995. Since joining NTT Information and Communication Systems Laboratories in 1995, he has been engaged in R&D of high-speed communication processing boards, reliable multicast delivery systems, secure electronic voting systems, and secure file delivery systems. From 2005 to 2008, he worked in a department supporting the development of systems in NTT WEST. He is currently engaged in R&D of cloud computing technology. As a result of organizational changes in April 2012, he is now with NTT Software Innovation Center.