Next-generation Data Hub for Secure and Convenient Data Utilization across Organizational Boundaries

Kei Ohmura, Hongjie Zhai, Shoko Katayama, Sakiko Kawai, Keiichiro Kashiwagi, Kenji Umakoshi, Yukiko Yosuke, and Tatsuro Kimura

Abstract

In a data-driven society, where balancing both economic development and solving social issues is anticipated, utilizing a variety of data across the boundaries of companies and organizations will be necessary. However, there are many challenges related to the handling of sensitive data and algorithms and acquiring desired data from diverse data sets in different companies and organizations. Data utilization across organization boundaries thus has not been widely carried out. In this article, we introduce our next-generation data hub and its key technologies for addressing these challenges to allow data to be used securely and conveniently across organizations.

Keywords: data sharing, data hub, data sandbox

1. What is a next-generation data hub?

As seen in smart cities and cross-enterprise digital transformation, efforts have begun to create new value and solve social issues through data sharing and utilization across companies and organizations. However, the following challenges must be addressed to expand such efforts not just to a limited part of society but to all of society.

- It is difficult for data users to find data that match the desired purpose from among the vast amount of data gathered and managed respectively by each company; it is also difficult for them to acquire such data quickly when needed.
- It is difficult for data providers to appropriately control the use of their valuable data. It is also difficult for them to understand the range in which their data are distributed and the track record of their use.
- Data or data-analysis-algorithm providers are also especially concerned about leakage of sensitive information due to their use for non-intended purposes when other companies are allowed to use data providers' sensitive data or data-analysis algorithms containing valuable information.

To address these challenges, we are engaged in the research and development of a next-generation data hub. This data hub is a data-sharing infrastructure that maintains governance of data by data providers while allowing data users to quickly and efficiently obtain necessary data kept at multiple locations. It allows data to be used securely and conveniently across the boundaries of companies and organizations.

There are three main components of our next-generation data hub (**Fig. 1**):

(1) Virtual data lake: It enables efficient data search and acquisition by virtually integrating data kept in multiple companies and organizations.

(2) Data broker: It enables efficient sending and receiving of data between multiple locations.

(3) Data sandbox: It enables execution of data and

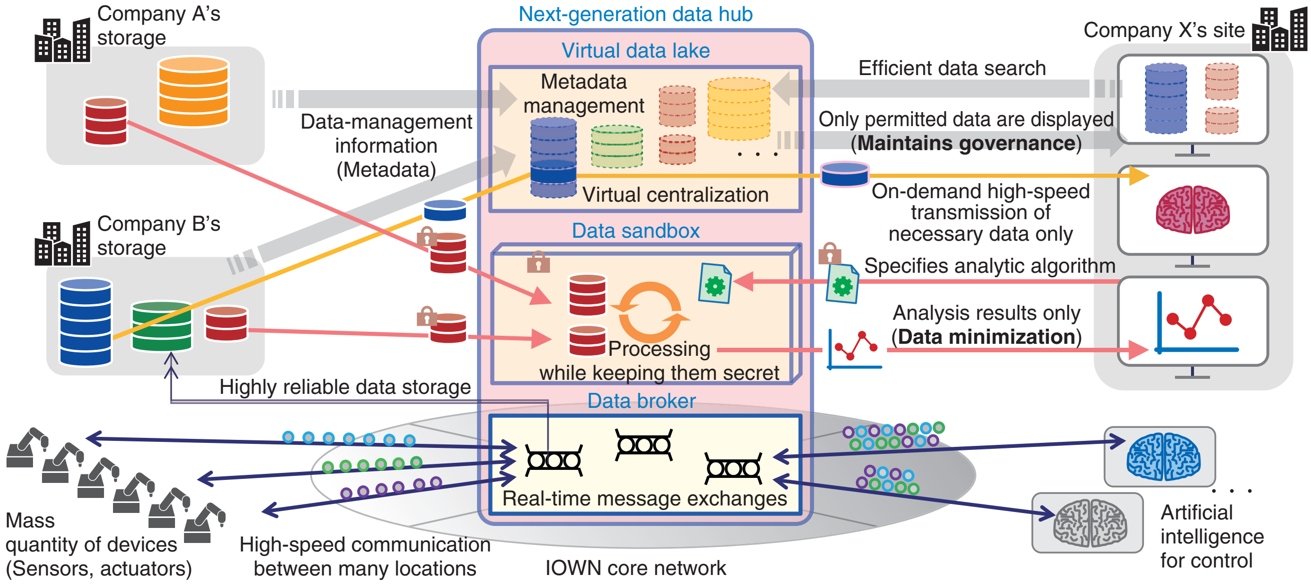
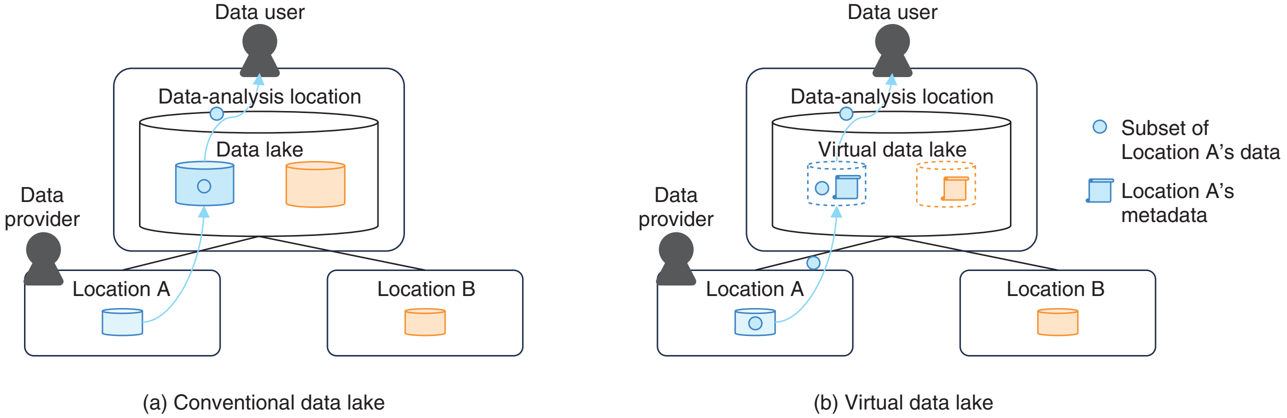

Company A's storage
Next-generation data hub
Virtual data lake
Metadata management
Data-management information (Metadata)
Virtual centralization
Company X's site
Efficient data search
Only permitted data are displayed (Maintains governance)
On-demand high-speed transmission of necessary data only
Company B's storage
Data sandbox
Specifies analytic algorithm
Processing while keeping them secret
Analysis results only (Data minimization)
Highly reliable data storage
Data broker
Real-time message exchanges
Mass quantity of devices (Sensors, actuators)
High-speed communication between many locations
IOWN core network
Artificial intelligence for control

Fig. 1. Next-generation data hub.



(a) Conventional data lake

(b) Virtual data lake

Fig. 2. Data-sharing models.

algorithms between companies while keeping them secret from each other.

We explain the problems to be solved with each component and their solutions in the following sections.

2. Virtual data lake

When using data that exist in multiple organizations and locations with the conventional model (as shown in **Fig. 2(a)**), data generated in each location are aggregated into a single location to create a vast data lake. Each user accesses the lake to use the data. However, this model has general problems such as the need to copy entire data sets, even when data users actually use just a small portion of the data. In addition, data providers face the problem of losing data governance because a large number of copies of their data are generated. It is thus difficult to conduct various analyses by sharing data between companies or organizations.

We are thus engaged in the research and development of a data-infrastructure technology called virtual data lake. This technology enables the data in various locations to be used without the need to aggregate them into a single location.

As shown in **Fig. 2(b)**, a virtual data lake does not collect data. Instead, it virtually aggregates and centralizes data by collecting their metadata[*1]. This makes it possible for data users to efficiently obtain the necessary data on demand for analysis and processing. For data providers, this technology makes it easy to maintain governance of their data by allowing them to always manage master data at their locations and providing data to data users upon request via the virtual data lake.

We are implementing this data lake from two perspectives: (i) data discovery and utilization, which is related to how to efficiently discover and use data necessary to meet data users' purposes from vast amounts of data with different formats and quality due to different acquisition histories, and (ii) data management and delivery, which asks how to efficiently manage ubiquitous data generated and growing daily in different locations, and how to deliver data to data users in an always usable condition and at the appropriate time. Next, we explain our efforts to address challenges associated with these two perspectives.

2.1 Data discovery and utilization

To enable data users to discover data they need from a vast amount of data, we assign and manage metadata, which explain data in detail in a unified, rule-based manner.

Some types of metadata, such as the semantic information of data, are assigned in advance by the data provider, while other types, such as format or quality information of data, are assigned automatically. By using metadata, data users can search data flexibly with various conditions. This makes it easy for them to narrow down to the data they need for achieving their purposes.

Metadata on the relationships between different data or the provenance of data—information on origin, distribution route, and processing executed for them—are also assigned and managed. Metadata on relationships between data make it easier for data users to find the data they need by traversing relevant data, even if they have just a vague clue about what to look for. Metadata on provenance of data make it possible to confirm, for example, that the data have not been improperly processed or that they have an unknown or suspicious origin. This enables data users to determine the data's reliability.

2.2 Data management and delivery

Managing dispersed data in their locations efficiently without aggregating them and enabling them to be obtained on demand presents several challenges. We first present our efforts to efficiently and remotely understand and manage the latest statuses of data, which are generated and updated daily at various locations. We then present our efforts to improve response time from data request to receipt so that it approaches that of a single data lake in which data are aggregated.

To allow data at all locations to be available to data users at any time, information on events in each location, such as data creation, update, and deletion, are collected with low latency and managed as metadata. By using up-to-date metadata, data users can handle the data as if they exist locally. For example, they can view a list of latest file information and issue a request to retrieve a file's content.

To ensure that the response time from a data request by a data user to its return is not so long as to present a practical use problem, we will also deploy a data format that carries out incremental data delivery and a caching mechanism within the virtual data lake. These solutions make it possible to suppress an increase in processing time due to data transmission in processes that repeatedly use the same data such as machine learning. We are also developing a prefetch function that dynamically caches data on the basis of the data's access pattern and frequency and a function for push-type distribution of data from each data location to the virtual data lake. We are designing and developing functions so data users and providers can plug in their algorithms into the virtual data lake according to their requirements.

3. Data broker

With the development of smart factories and connected vehicles, large amounts of sensors and terminals are required to be connected in one single network, where messages are exchanged for monitoring and control of factories or vehicles. For accurate monitoring and control, it is necessary to collect data from terminals and send feedback to terminals reliably in an extremely short amount of time. Therefore, we are developing a new broker technology to enable low-latency and high-reliability exchanging of

[*1] Metadata: In this article, metadata refer to "data for explaining data." Such metadata include information about the data's location, creator, and format.

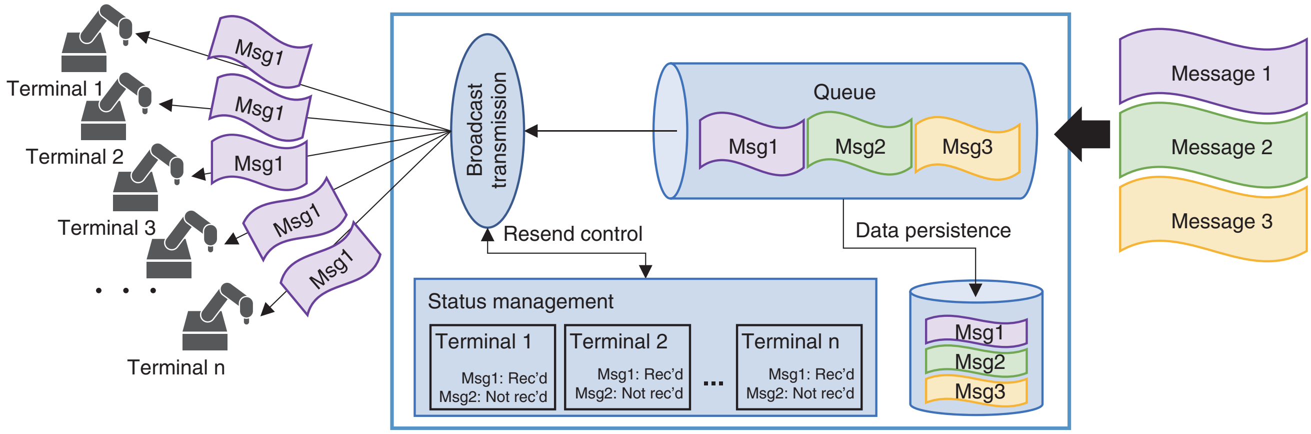


Fig. 3. Persistence and resend mechanism.

messages among huge quantity of terminals.

Conventional broker technologies prioritize one of two aspects: (i) efficient transmission of the same information to a large number of terminals, in which case reliability is sacrificed as message persistence is not ensured and resending of messages is not carried out, or (ii) reliable transmission, in which case transmission to a large number of terminals cannot be supported. It is difficult to support requirements for both aspects at the same time. To achieve both sets of requirements, a challenge is the resend process for messages that are not received by terminals. To resend messages, it is necessary to record the transmission status of every message for each terminal, and managing statuses becomes extremely costly when a large number of terminals are connected in a network. As a result, transmission delays occur due to insufficient computing resources of brokers. We are working to address this challenge by improving transmission protocols and status-management algorithms (**Fig. 3**).

By accumulating these efforts, we will achieve a new data broker that allows low-latency, high-reliability message exchanges between a large number of terminals.

4. Data sandbox

4.1 Background

As stated above, despite the expectation that new value will be created by analyzing sensitive data with an algorithm held by different companies, companies have not collaborated much with one another due to concerns about leakage of their data and algorithms.

A simple method for analyzing data with an algorithm without disclosing data or the algorithm to each other is to leave data and the algorithm to a third party (platform provider). The platform provider is asked to return only the analyzed result. With this method, there is no need for companies to disclose data and an algorithm to one another. However, the remaining issue is that data and an algorithm are disclosed to the platform provider.

We are researching and developing a data-sandbox technology that, while on the basis of the model of the platform provider performing computation on behalf of data providers and algorithm providers, enables the analysis of data with an algorithm to be kept secret from the platform provider. We seek to achieve utilization of data and algorithms such as the following with this technology.

- The ability for competing companies to bring together data, process them, and share the results among themselves. None of the companies disclose their own data and algorithms to others or the platform provider.
- The ability for a company to analyze its valuable data by using a secret analysis program developed by another company and obtain analytic results. None of the companies disclose original data or analytic programs to each other or to the platform provider.

4.2 Technological challenges and approaches for solutions

To enable data utilization as described above, the following issues must be addressed.

- Issue 1: Unauthorized execution of algorithms

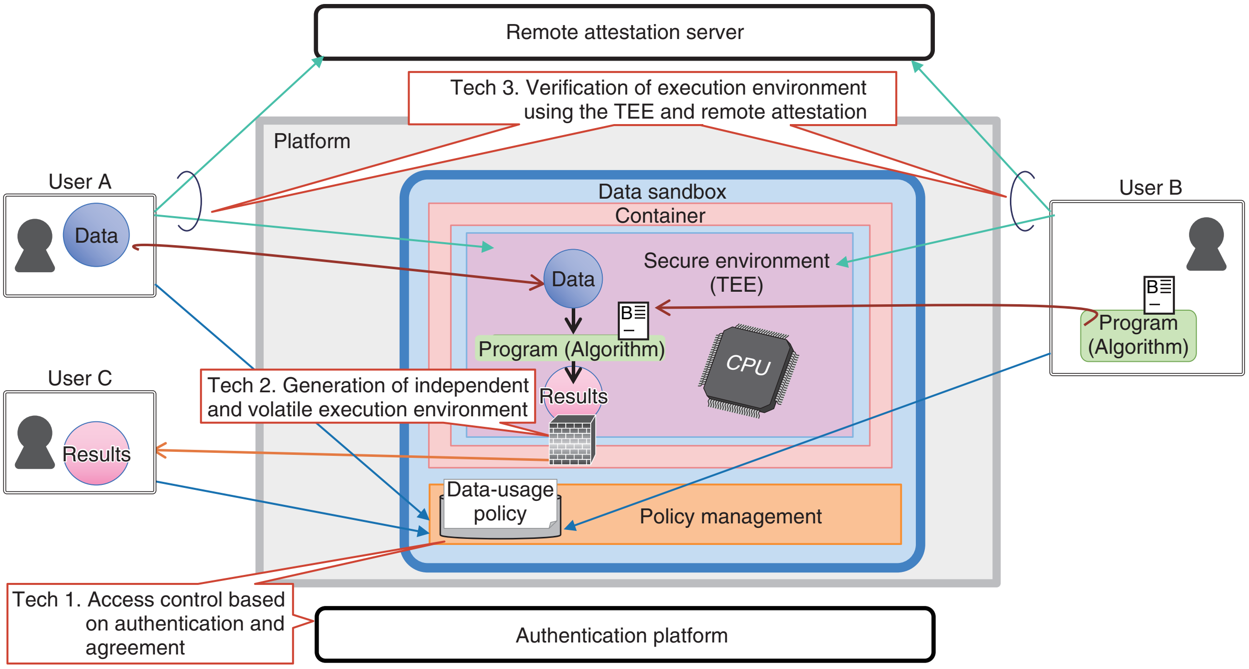

Remote attestation server
Tech 3. Verification of execution environment using the TEE and remote attestation
Platform
User A
Data
Data sandbox
Container
Secure environment (TEE)
Data
Program (Algorithm)
Results
CPU
User B
Program (Algorithm)
User C
Results
Tech 2. Generation of independent and volatile execution environment
Data-usage policy
Policy management
Tech 1. Access control based on authentication and agreement
Authentication platform

Fig. 4. Data-sandbox technology.

by spoofing, which results in the spoofer illegally acquiring the original data and results

- Issue 2: Leakage of original data and results due to execution of faulty algorithms
- Issue 3: Unauthorized acquisition of original data and algorithms by the platform provider

Our data sandbox addresses these issues by combining the following technologies (**Fig. 4**).

Technology 1: Authentication and agreement-based access control

The data sandbox authenticates users (data and algorithm providers and users of result data) to prevent spoofing. It also controls access to original data, algorithms, and result data in accordance with agreements (data-usage policy) set in advance by users to prevent unauthorized execution of algorithms and acquisition of data.

Technology 2: Creation of volatile and independent execution environment

The data sandbox creates an independent execution environment for each data-usage policy agreed upon by users. Access to the outside from this execution environment is also restricted. This prevents algorithms from taking data to the outside in violation of the data-usage policy. Furthermore, the execution environment is deleted after processing of an algorithm is completed, preventing the leakage of data and algorithms.

Technology 3: Verification of the execution environment by applying TEE and remote attestation

The Trusted Execution Environment (TEE)[*2] and remote attestation[*3] are technologies for preventing operators and administrators of the execution environment, such as the platform provider, from accessing data and algorithms. By applying these mechanisms, the data sandbox allows users themselves to verify that the data and algorithms set into the execution environment are those agreed upon by the data-usage policy and that the execution environment being used is generated using the TEE, with the data and algorithms being kept secret. Applying these mechanisms prevents the platform provider from

[*2] TEE: Trusted Execution Environment (TEE) is an isolated execution environment architected such that the memory region is encrypted by the central processing unit (CPU) so that administrative users of an operating system cannot read the content of the memory. This mechanism has been used mainly in mobile terminals and embedded devices. However, it has been incorporated into many server CPUs by manufacturers such as Intel and AMD.

[*3] Remote attestation: Function provided by CPU vendors as a method for users to confirm the authenticity of the TEE. By obtaining information about the TEE's configuration and having the CPU vendor attest to its authenticity remotely via the Internet, users can verify that the TEE has been created using the genuine functions provided by the CPU vendor and it was not falsified.

accessing data and algorithms while allowing users to obtain results that come from analyzing data with an algorithm.

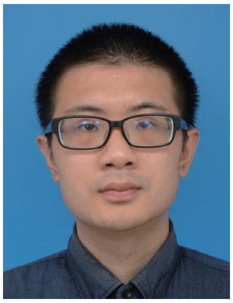
5. Going forward

In this article, we introduced the main components of our next-generation data hub that we are researching and implementing: virtual data lake, data broker, and data sandbox. Our next-generation data hub will enable safer, more secure, and more efficient data sharing. It will thus foster the creation of new value and solutions for social issues through the mutual use of highly confidential data and algorithms across companies and organizations, which had been considered difficult to date. We will further accelerate the research and development of these elemental technologies and evaluate them with partners to contribute to achieving a data-driven society as early as possible.

Kei Ohmura

Senior Research Engineer, Data Sharing Infrastructure Project, NTT Software Innovation Center.

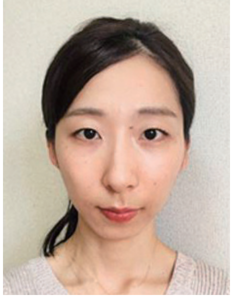
He received an M.E. from Waseda University, Tokyo, in 2009. Since joining NTT the same year, he has been engaged in developing platforms for cloud, Internet of Things (IoT), and artificial intelligence.



Hongjie Zhai

Engineer, Data Sharing Infrastructure Project, NTT Software Innovation Center.

He joined NTT after he received a Ph.D. in computer science and information technology at Hokkaido University in 2018 and is currently focusing on researching a big-data platform.



Shoko Katayama

Research Engineer, Data Sharing Infrastructure Project, NTT Software Innovation Center.

She received an M.E. from Waseda University, Tokyo, in 2015 and joined NTT the same year. She is currently focusing on a data-sharing infrastructure.

Sakiko Kawai

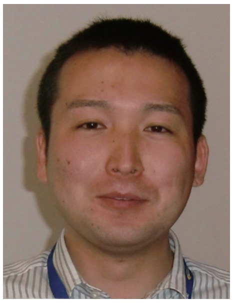
Research Engineer, Data Sharing Infrastructure Project, NTT Software Innovation Center.

She received an M.E. from Tokyo Metropolitan University in 2013 and joined NTT the same year. She is currently focusing on data-sharing and data-management technology.

Keiichiro Kashiwagi

Senior Research Engineer, Data Sharing Infrastructure Project, NTT Software Innovation Center.

He received an M.E. in information engineering from Waseda University, Tokyo, in 2008 and joined NTT the same year. He received a Ph.D. in pure and applied mathematics from Waseda University in 2013. He is currently focusing on secure data-sharing technologies. He is a member of the Information Processing Society of Japan (IPSJ).



Kenji Umakoshi

Senior Research Engineer, Social Information Sharing Research Project, NTT Social Informatics Laboratories.

He received an M.E. from Waseda University, Tokyo, in 2009 and joined NTT the same year. He has been researching and developing ubiquitous/IoT computing, smart room/factory, and data-sharing platforms. He also worked as a product manager of a cloud service at NTT Communications Corporation from 2014 to 2016.

Yukiko Yosuke

Senior Research Engineer, Data Sharing Infrastructure Project, NTT Software Innovation Center.

She received an M.S. in mathematics from Osaka University in 1997 and joined NTT the same year. She is currently focusing on a data-sharing infrastructure. She is a member of IPSJ.

Tatsuro Kimura

Senior Research Engineer, Data Sharing Infrastructure Project, NTT Software Innovation Center.

He received an M.S. in agricultural and life science from the University of Tokyo in 2001 and joined NTT the same year. He is currently engaged in researching data hubs and data lakes.