iChie: Speeding up Data Collaboration between Companies

Naoto Yamamoto, Daisuke Tokunaga, and Seiichiro Mochida

Abstract

At NTT Software Innovation Center, we are researching and developing iChie, which is a technology for the virtual integration of disparate databases into a single entity to promote the creation of new value of data linkages between organizations and companies. This article discusses the issues of inter-company data linkage and the technical features of iChie for resolving these issues.

Keywords: federated database system, data virtualization, privacy policy

1. Expectations and reality in inter-company data linkage

There has been growing interest in carrying out data linkage across the boundaries between companies and industries. For example, if the sales histories of retailers are combined with the human-flow data collected by public transportation networks, it should be possible to analyze trade areas more precisely and make flow design more efficient. It should also be possible to improve product traceability by integrating the sales histories of wholesalers and retailers with the assembly histories of manufacturers and manufacturing histories of suppliers.

However, when this type of inter-company data analysis is conducted, the data must be gathered in a single location (where the data analysis is to be conducted). To protect personal information and trade secrets, these data must also be anonymized and/or concealed in some way. This increases the granularity of the data and reduces their value.

Even within the same company, there are many cases in which different departments create and operate their own databases, giving rise to issues similar to those of inter-company data linkage. Some companies have taken the approach of temporarily storing the data from all their databases in a data lake then reconfiguring the data as data marts for specific analysis targets. When using a data lake that can store data in any form, either structured or unstructured, it is possible to store and use data collected from each database at a single centralized location. However, there are high barriers to the collection of sensitive data, such as personal information and trade secrets, in an external data lake. For this reason, iChie—a technology for the virtual integration of disparate databases into a single entity to promote the creation of new value of data linkages between organizations and companies—adopts an approach called *data virtualization* whereby the distributed databases of individual companies are left intact and provided only as centralized endpoints for analysis by presenting them as virtualized databases.

Table 1 summarizes the expectations and reality of inter-company data linkage.

iChie provides two functions:

1) Data-transfer control based on network characteristics
2) Mandatory application of privacy policy to data users by data owners (privacy-policy enforcement)

Data-transfer control can speed up data transfer, and privacy-policy enforcement can overcome the high barriers to the collection of sensitive data. The following sections provide details of these features.

Table 1. Expectations vs. reality in inter-company data linkage.

	Expectation	Reality	Benefits offered with iChie
1.	Supports real-time on-demand data linkage to remote databases (DBs).	Data transfers are time-consuming and make real-time, on-demand linkage impossible.	• Reduces the volume of data transfers by transferring small quantities of data to locations where there are large quantities of data. • Selects the optimal data transfer route based on the network quality between DBs.
2.	Results can be returned from external data analysis while protecting data privacy and trade secrets.	When data are handled externally, they must be protected through processing such as anonymization or concealment, which reduces the benefit of analyzing these data.	• If a specified data item is one that the data owner wishes to prohibit from being transferred externally, then it is possible to conduct analysis without transferring this item externally.

2. Data-transfer control based on network characteristics

Most companies use applications such as business intelligence (BI) tools to obtain useful information for decision making by aggregating and analyzing large amounts of data. When BI tools collect data, they submit Structured Query Language (SQL) queries to databases, which transfer data in response to these queries. In almost all cases, inter-company data linkage involves databases scattered throughout different geographical locations and on different networks. When a BI tool is used to collect data from each database in such situations, the quality of the networks connecting it to these databases becomes a bottleneck, requiring a long time for data transfer to complete. BI involves searching for correct answers while changing the combination of data through trial and error. Therefore, long data-transfer times lead directly to a reduction in the number of trials, resulting in lower analysis quality.

With iChie, this issue is resolved by taking two approaches:

The first is to reduce the volume of data transfers by transferring smaller quantities of data to places where larger amounts of data are stored. When using a JOIN query to join tables from multiple databases, the data are collected and combined at a (single) location where the SQL queries are issued. With iChie, on the other hand, statistical information is used to compare the data sizes corresponding to hits in partial queries submitted to the tables to be joined. Instead of collecting hit data where the SQL query was submitted, the data are sent from the database(s) yielding fewer hits to one yielding more hits. The data are then joined, and the results are returned to the location where the SQL query was submitted. Since the joined table represents an intersection of the original tables, it contains less data than the original large tables. **Figure 1** shows an example where two databases are joined. When JOIN operations are carried out across three or more databases to minimize the total amount of data transferred, an execution plan is devised to determine which database's data should be sent to what other database and in what order.

The second approach is to select the optimal data-transfer route based on network quality. As shown in Fig. 1, there are multiple possible data-transfer paths when transferring data between databases or when sending query responses to BI tools. With iChie, the results of effective bandwidth measurements obtained when previously transferring data over these routes can be fed back to the query-execution plan, making it possible to select the optimal data-transfer paths (**Fig. 2**).

3. Mandatory application of privacy policy to data users by data owners

iChie includes functions that support integrated analysis with privacy in mind. When inter-company data linkage is used to handle data, such as personal information or trade secrets that must be managed more securely, processing, such as anonymization and concealment, must be carried out to protect such data. However, excessive anonymization can make the data too granular and unsuitable for analysis.

For this reason, iChie uses a mechanism called privacy-policy enforcement to provide functions for analyzing data without exposing them to the outside. This mechanism forces data users to apply the concealment/anonymization policy set by the data owner (database administrator).

For example, suppose a shopping mall collects

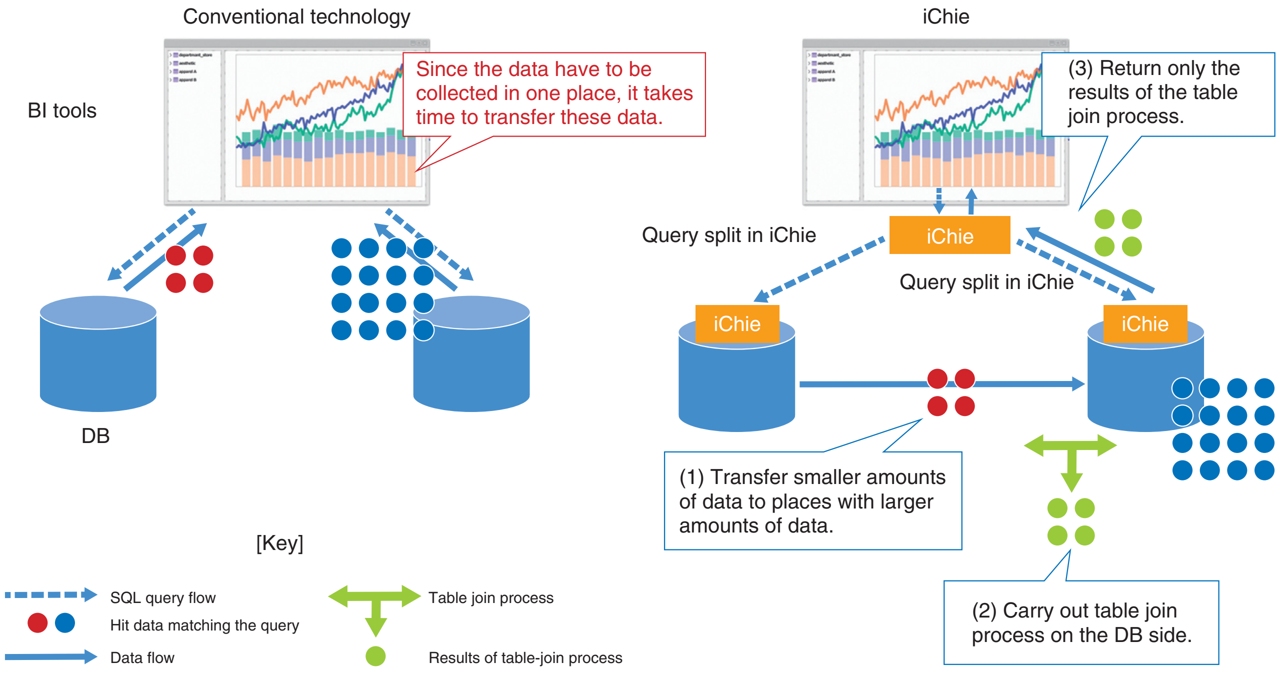

Conventional technology
iChie
BI tools
Since the data have to be collected in one place, it takes time to transfer these data.
(3) Return only the results of the table join process.
Query split in iChie
iChie
Query split in iChie
iChie
iChie
DB
(1) Transfer smaller amounts of data to places with larger amounts of data.
(2) Carry out table join process on the DB side.
[Key]
SQL query flow
Hit data matching the query
Data flow
Table join process
Results of table-join process

Fig. 1. Mechanism for reducing the volume of data transfers.

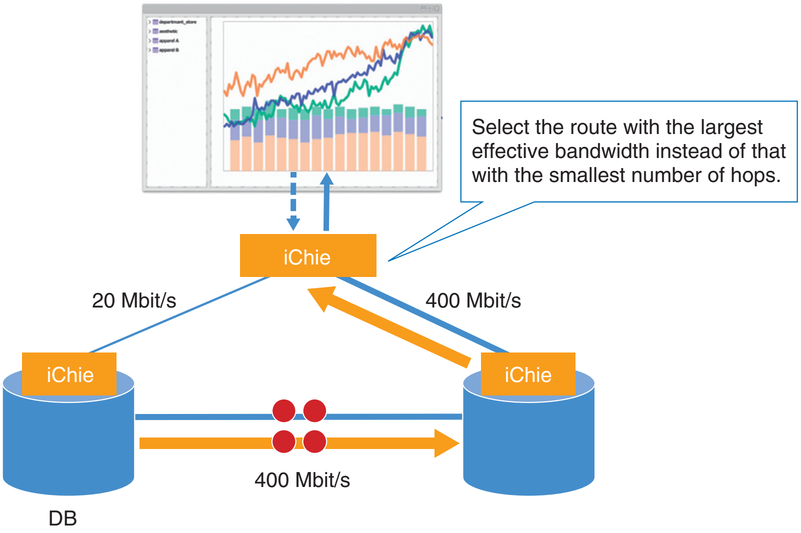
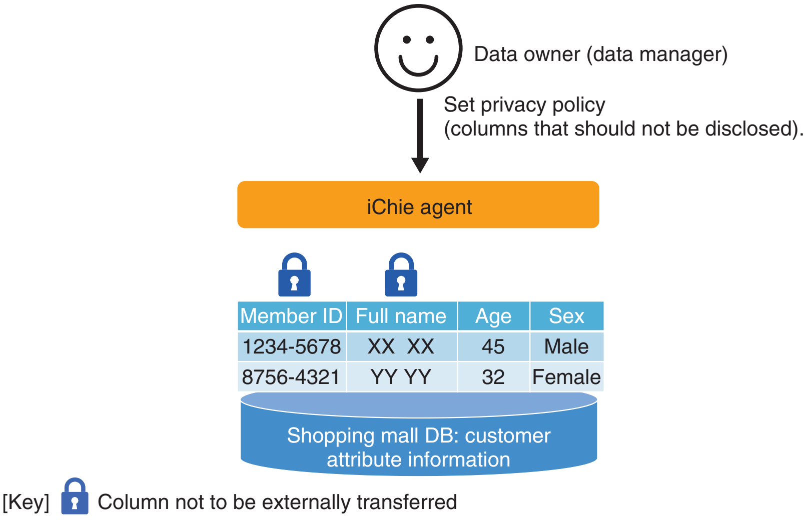

Select the route with the largest effective bandwidth instead of that with the smallest number of hops.
iChie
20 Mbit/s
400 Mbit/s
iChie
iChie
400 Mbit/s
DB

Fig. 2. Route selection between DBs.

customer-attribute information while one of its tenants collects purchase-history information. The shopping mall's security policy dictates that member identifiers (IDs) and names must not be disclosed. In this case, the shopping mall (data owner) sets a privacy policy in advance for the iChie agent to prevent the disclosing of member IDs and names (**Fig. 3**). Now suppose a BI tool user wants to check the purchase date and purchaser age and sex for each purchased product. To check this information, it is necessary to carry out a JOIN operation on query responses obtained from the shopping mall and tenant databases.


Data owner (data manager)
Set privacy policy
(columns that should not be disclosed).
iChie agent
Shopping mall DB: customer attribute information
[Key]
Column not to be externally transferred

Member ID	Full name	Age	Sex
1234-5678	XX XX	45	Male
8756-4321	YY YY	32	Female

Fig. 3. Example of a data owner's privacy policy settings.

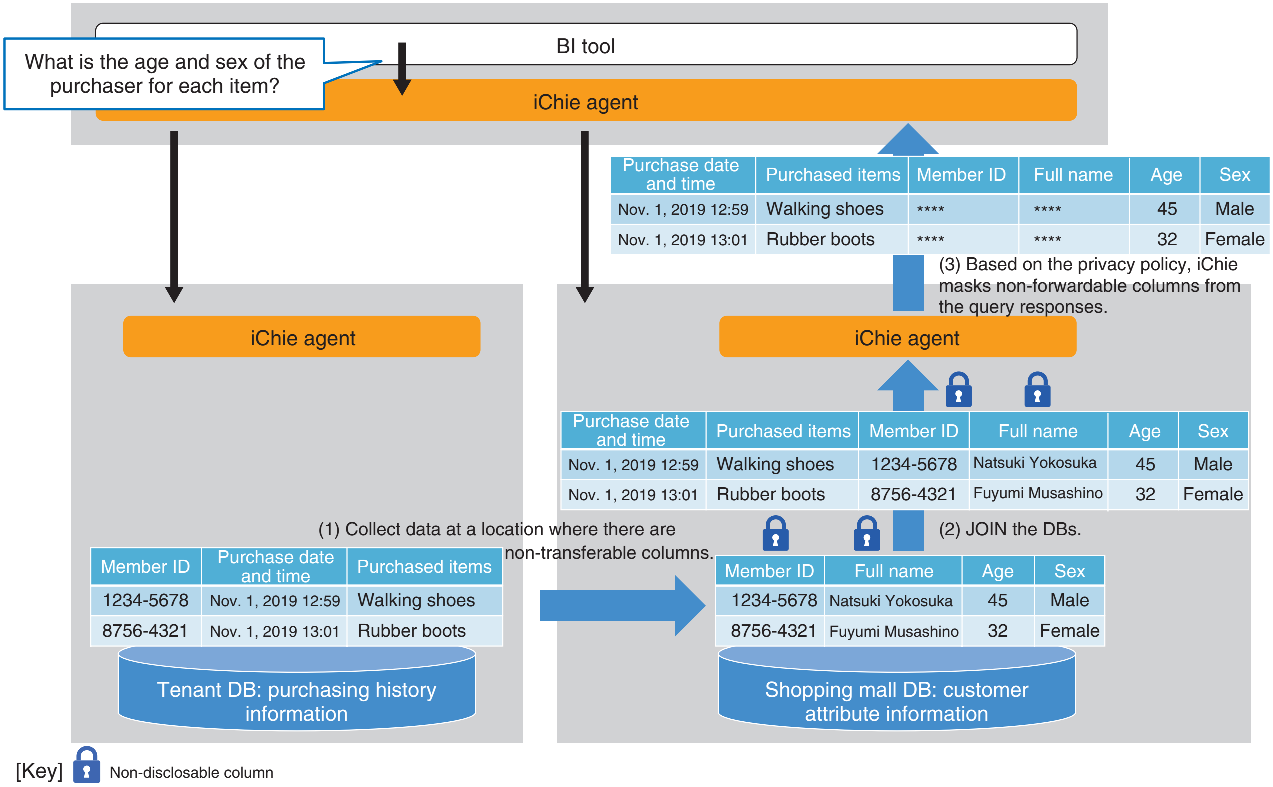
In this case, iChie creates a query-execution plan for each database so that the JOIN operation including the columns subject to the privacy policy is carried out in the shopping mall database. From the results of this JOIN operation, the columns that are not to be disclosed (i.e., member IDs and names) are masked and sent back to the BI tool. Therefore, data analysis can be conducted without divulging confidential data (**Fig. 4**).

4. Future prospects

At the 2011 annual meeting of the World Economic Forum in Davos, Switzerland, it was reported that personal information will become the *new oil* (i.e., a valuable resource) in the 21st century. Although data are as valuable as oil, the value of data is maximized by combining data from different sources and conducting appropriate analysis. Therefore, the need for data linkage between companies and industries is expected to increase.

Inter-company data linkage involves many other issues besides those mentioned in this article. For example, in data linkage between companies, it is seldom the case that the same data are stored using the same column names and data types, so when joining tables from different companies, it is necessary to sort by what each column name refers to. We are therefore investigating a technique for iChie whereby the semantic structure of data in disparate databases can be analyzed and converted into a common representation format. We are also exploring the development of a technique that uses collaborative distributed machine learning to eliminate the need for companies to share trade secrets in the form of real data. Instead, they would use their own data to create a learning model that can be shared and integrated with other models to achieve the same effect as that of sharing real data.

At NTT Software Innovation Center, we will promote real-world applications by collaborating not only with the NTT Group but also with various other partners.


BI tool
What is the age and sex of the purchaser for each item?
iChie agent
iChie agent
iChie agent
(3) Based on the privacy policy, iChie masks non-forwardable columns from the query responses.
(2) JOIN the DBs.
(1) Collect data at a location where there are non-transferable columns.
Tenant DB: purchasing history information
Shopping mall DB: customer attribute information
[Key] 🔒 Non-disclosable column

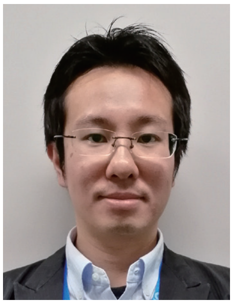
Purchase date and time	Purchased items	Member ID	Full name	Age	Sex
Nov. 1, 2019 12:59	Walking shoes	****	****	45	Male
Nov. 1, 2019 13:01	Rubber boots	****	****	32	Female

Purchase date and time	Purchased items	Member ID	Full name	Age	Sex
Nov. 1, 2019 12:59	Walking shoes	1234-5678	Natsuki Yokosuka	45	Male
Nov. 1, 2019 13:01	Rubber boots	8756-4321	Fuyumi Musashino	32	Female

Member ID	Purchase date and time	Purchased items
1234-5678	Nov. 1, 2019 12:59	Walking shoes
8756-4321	Nov. 1, 2019 13:01	Rubber boots

Member ID	Full name	Age	Sex
1234-5678	Natsuki Yokosuka	45	Male
8756-4321	Fuyumi Musashino	32	Female

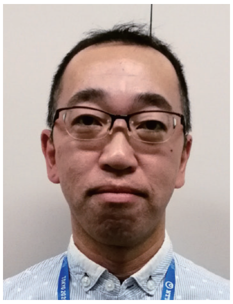
Fig. 4. Example of privacy policy enforcement.



Naoto Yamamoto

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

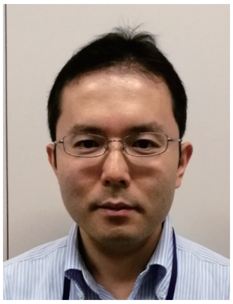
He received a B.A. and M.M.G. in bioinformatics from Keio University, Tokyo, in 2004 and 2006. He joined NTT Information Sharing Platform Laboratories in 2006, where he was involved in research and development (R&D) of e-payment. He moved to NTT WEST in 2009, where he worked on domain name systems and software-defined networking. He transferred to NTT Software Innovation Center in 2014. His current research interests are in the system architecture of the Internet of Things (IoT) management platform.



Daisuke Tokunaga

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

He received a B.E. and M.E in information engineering from Kyushu Institute of Technology, Fukuoka, in 1997 and 1999. He joined NTT Information Sharing Platform Laboratories in 1999, where he was involved in R&D of asynchronous transfer mode networks, public wireless network authentication service for mobile devices, virtual computing and networking infrastructure management, and software engineering on network carrier services. He moved to NTT WEST in 2012, where he worked on software service development for NTT Next Generation Network. He transferred to NTT Software Innovation Center in 2015. His current research interests are in the system architecture of the IoT management platform.



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 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 IP (Internet protocol) phone services. He transferred to NTT Software Innovation Center in 2012. His current research interests are in the system architecture of the IoT management platform.