Improving Reliability of Data Distribution Across Categories of Business and Industries with Chain Data Lineage

● Masazumi Matsubara ● Takeshi Miyamae ● Akira Ito ● Ken Kamakura

Today, companies are overflowing with a variety of data, inside and out. Recently, there is a growing movement to push the co-creation of innovative digital businesses by distributing and utilizing valuable data that cannot be obtained independently between different companies. The realization of a world like this, however, requires the data shared between companies to be reliable. Specifically, it must be possible to verify where the data have come from and how they have been processed and, when personal data are included, whether or not individuals have given consent for the data to be provided. Fujitsu Laboratories is moving ahead with research and development relating to data distribution and utilization technology that resolves these issues and enables the safe use of data. This paper first presents the issues relating to data reliability in data distribution. Then, it describes the Fujitsu-developed Chain Data Lineage technology, which improves the reliability of data exchanged across different categories of business and industries, together with examples of its application.

1. Introduction

Today, companies are overflowing with a variety of data, inside and out. For example, information such as customer master data, retail point of sales (POS) system data, and the product purchase histories on electronic commerce (EC) websites have long been regarded as important. But recently, the types and amount of data that can be used by businesses have been expanding more and more. These include IoT data such as monitoring sensor data used in factory production lines; open data such as weather information and demographic surveys; and personal data such as automobile driving data that indicate the behavior of individuals.[1)]

Companies are driving digital innovation through the effective utilization of these types of data. In other words, by combining data and using AI and data analysis to find new value, new businesses that were previously not possible can be created, and conventional business models can be radically transformed.

Another trend in data utilization in recent years is toward extending it to the sharing of data between companies. Data whose value cannot be exploited by a company itself can be shared with other companies to “co-create” innovative digital businesses that were not previously possible.

Fujitsu Laboratories aims to create a world where such data-driven digital co-creation can be realized.[2)] However, the reliable handling of data is an important precondition to creating this kind of world. To achieve this, Fujitsu Laboratories is conducting research and development relating to data distribution and utilization technology to enable the safe use of data.

This paper first presents the issues relating to data reliability in data distribution. Then, it describes the Fujitsu-developed Chain Data Lineage technology, which improves the reliability of data exchanged across different categories of business and industries, together with examples of its application.

2. Issues for achieving highly reliable data distribution

When distributing data between companies, the target data is exchanged based on a contract between the companies owning the data that was generated from human activities and businesses, and the companies that will receive and utilize this data. Companies that receive the data use the data as it is or combine them with other data to create AI models or perform analysis, which is then used in the companies' own businesses. Moreover, new businesses can grow by distributing this newly generated data to third-party companies.

However, if low-quality, unreliable data is used at this time, it will lead to issues such as lowering the quality of the provided services. Various laws have been enacted recently to protect personal information, most notably the EU General Data Protection Regulation (GDPR). As a result of these laws, when personal data including personal information is provided to third-party companies, the consent of the individual must be obtained in advance. Therefore, it is necessary to devise methods for the active utilization of data that also make it easy for individuals to understand how their personal data is being used.

In the future, the amount of diverse and useful data will only increase as most companies and organizations become involved in generating or processing personal data. On the other hand, there are concerns about the decline in reliability as data is exchanged and processed across a large and indefinite number of companies, which has increased the need to verify the origin of data.

Most data distributed between companies involves the secondary use of other data, with processing or analysis applied to this data, such as adding, deleting, or converting information. This means that in order to increase the reliability of data distribution, history information must be managed that describes what kind of data is used and how the data was processed or analyzed.

But in the past, history information relating to data processing or analysis was managed separately by each company and this information was not transmitted to the companies by which the data was received. Also, if there are changes such as in the personal data held by each company, the purpose of use of the data, the data provided among personal data, or the destination of the data, then consent must be obtained from individuals through the data provider again. Furthermore, the data user must obtain the personal data individually based on this consent. As a result, a lot of work is required to reconfirm the consent of individuals and obtain personal data individually.

3. Developed technologies

To address the issues of data reliability described in the previous section, Fujitsu Laboratories developed Chain Data Lineage[3] to improve the reliability of data exchanged across different categories of business and industries. These technologies are comprised of data history management technology and consent management technology (**Figure 1**). This section describes each technology in detail.

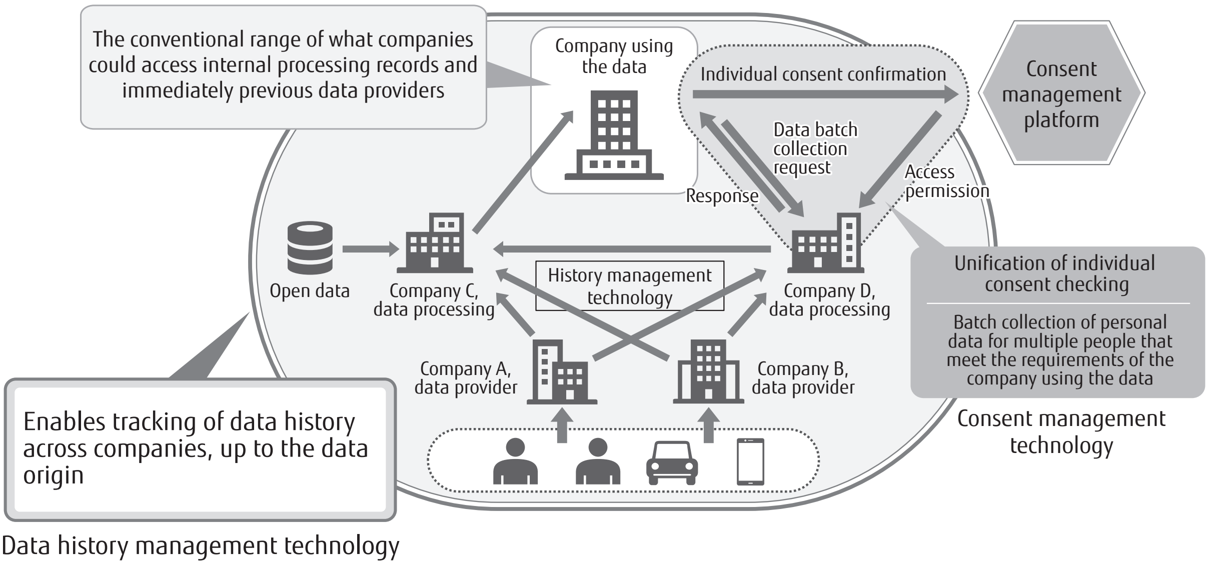
The conventional range of what companies could access internal processing records and immediately previous data providers
Company using the data
Individual consent confirmation
Consent management platform
Data batch collection request
Response
Access permission
Open data
Company C, data processing
History management technology
Company D, data processing
Unification of individual consent checking
Batch collection of personal data for multiple people that meet the requirements of the company using the data
Company A, data provider
Company B, data provider
Enables tracking of data history across companies, up to the data origin
Consent management technology
Data history management technology

Figure 1
Characteristics of Chain Data Lineage.

Note that Fujitsu plans to commercialize these technologies by implementing them in products such as the FUJITSU Intelligent Data Service Virtuora DX Data Distribution and Utilization Service[6]. This service is based on the Virtual Private digital eXchange (VPX) technology[4), 5)] that was developed by Fujitsu Laboratories by applying blockchain technology.

3.1 Data history management technology

Data history information in a data distribution system is a graph data in which data processing records or data trading records are connected in a time series. Chain Data Lineage first standardizes the formats of the data processing records in companies and the data trading records in the data trading system to enable interoperable management of all the data history information (**Figure 2**). To improve the tamper-resistance of data distribution history information, the data trading records are recorded on a blockchain.

Furthermore, we apply hash chain technology[note 1)], which is also used in blockchain technology, to all the history information including processing records. In other words, a hash value of each data record is embedded in each of next data records. This makes it easy to detect falsification of all the past history information, which increases the reliability of the history information. At the same time, the hash value of the data body is also embedded in each data record, which makes it easy to detect falsification or replacement of the data body and perform integrity verification[note 2)] of the distributed and processed data. This has been integrated into the history management technology.

These technologies provide the reliability required to guarantee the integrity of the distributed data and validity of the history information, which removes the need for third-party certification such as trusted third party (TTP) and greatly reduces system costs. As an ordinary use case history information can be attached to each data body to enable verification of the data on the device of the end user.

In Chain Data Lineage, data trading records, which require high security, are managed on a blockchain. On the other hand, data processing records, which are generated relatively often, are managed in databases outside the blockchain by each company. This achieves a good balance between security and scalability. **Figure 3** shows a diagram of the system configuration when sharing data history information

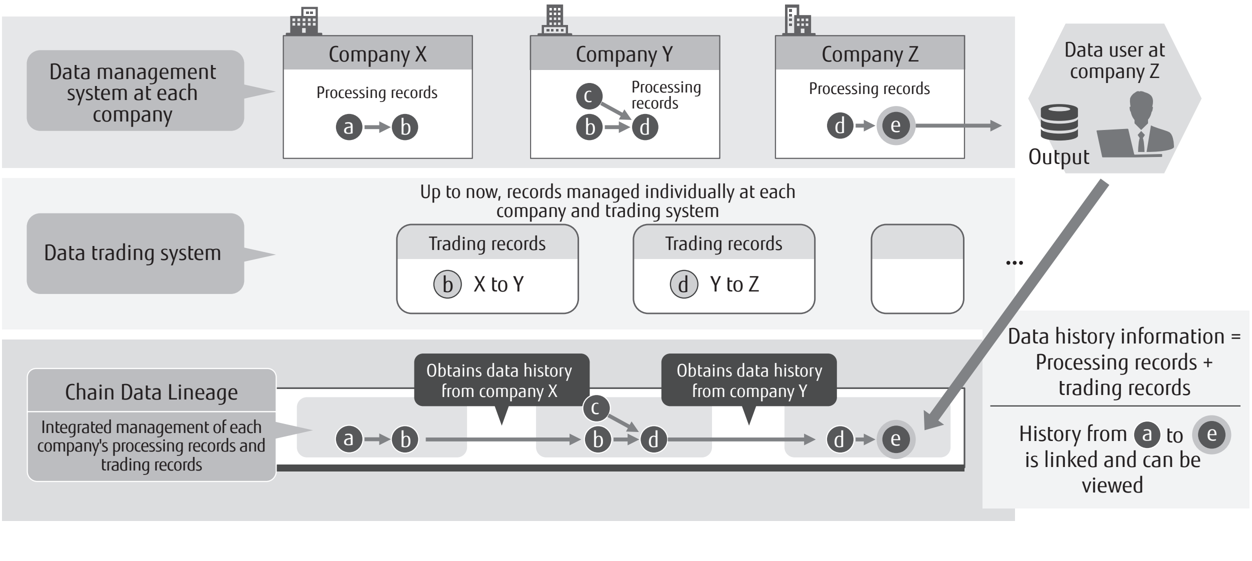
Data management system at each company
Company X
Processing records
a → b
Company Y
Processing records
c
b → d
Company Z
Processing records
d → e
Data user at company Z
Output
Up to now, records managed individually at each company and trading system
Data trading system
Trading records
b X to Y
Trading records
d Y to Z
...
Chain Data Lineage
Integrated management of each company's processing records and trading records
Obtains data history from company X
Obtains data history from company Y
a → b
c
b → d
d → e
Data history information = Processing records + trading records
History from a to e is linked and can be viewed

Figure 2
Data history management technology.

[note 1)] Technology that makes it difficult to falsify certain data by repeatedly applying special mathematical functions (hash functions) that satisfy certain characteristics such as unidirectionality.

[note 2)] The prevention of false entry, rewriting, deletion or mixing up of data, whether intentional or accidental, or the detection of such incidents.

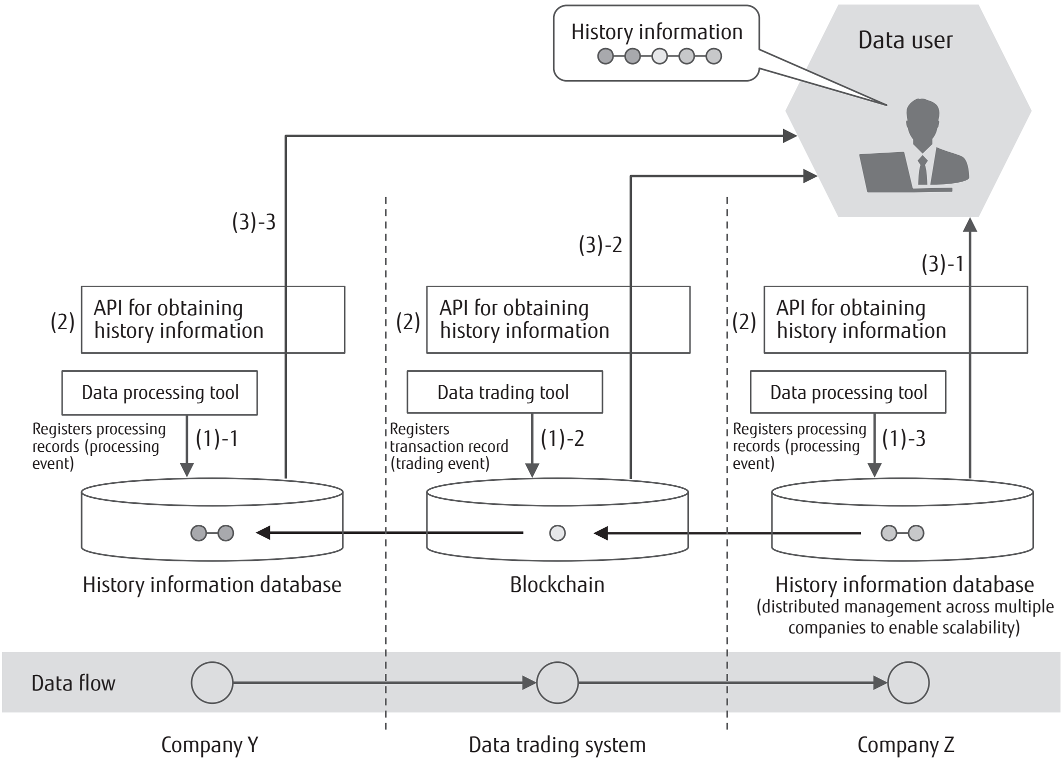
History information
Data user
(3)-3
(3)-2
(3)-1
(2) API for obtaining history information
(2) API for obtaining history information
(2) API for obtaining history information
Data processing tool
Data trading tool
Data processing tool
Registers processing records (processing event)
(1)-1
Registers transaction record (trading event)
(1)-2
Registers processing records (processing event)
(1)-3
History information database
Blockchain
History information database
(distributed management across multiple companies to enable scalability)
Data flow
Company Y
Data trading system
Company Z

Figure 3
Sharing data history information between companies.

between companies.

To record the data history information, each of data trading systems and data processing tools extracts, respectively, the trading and processing events and then records them in the history information database that they operate themselves, as shown in Figure 3 (1). The history information database is distributedly stored across multiple companies to improve scalability. An application programming interface (API) that obtains history information is provided by all the history information databases, and the history information managed by each company is disclosed to other companies participating in data trading, as shown in Figure 3 (2). A participant who wants to obtain a history information calls each company's API in sequence from a previous data record to farther ones, finally arriving at an original data record, as shown in Figure 3 (3).

Fujitsu has developed the VPX system, which uses blockchain technology for data trading records, and Apache Atlas, an open source software (OSS) metadata repository, which manages data processing records within a company, as Chain Data Lineage components and tested their interoperation.

A major obstacle to the spread of Chain Data Lineage as a method of implementing history management based on the hash chain is the difficulty in enabling the data trading and processing tools in each company to support the data hash chain format of the history information. To resolve this issue, in the future, the history information format and the APIs for registering and obtaining history information shall be standardized through organizations such as the Data Trading Alliance (DTA) in the case of Japan. This will improve interoperability to enable companies to create an environment where it is easy to generate history information in hash chain format. Methods such as demonstration tests in the Japanese Cabinet Office's Strategic Innovation Program (SIP) shall also be used to enlighten people on the importance of history information based on hash chain technology to improve the reliability of data distribution. Furthermore, Fujitsu will continue to encourage the Japanese Government to develop laws requiring each company to have storage for history information. From a technical perspective, we will provide libraries to easily generate the hash chain to lessen the development burdens for data trading and processing tools.

3.2. Consent management technology

Standards related to distributing personal data between companies include OAuth[7] for authorization and the User-Managed Access (UMA)[8] protocol based on this authorization. However, both of these standards were designed on the assumption that the data is held by the individuals themselves. This means that a new design is required when, for example, a medical facility wants to send all of its stored medical information for which it has already received consent to a different medical facility.

Figure 4 shows a typical configuration of OAuth and its UMA extension. The UMA protocol is a model provided to third parties that certifies data on the basis of the data owner. In this model, multiple communications are repeatedly generated between each server and client for as many users as are present, which generates a huge number of messages.

To resolve this problem, the UMA protocol has been extended to enable the batch acquisition of data for a large number of people. **Figure 5** shows the main processing sequence. By specifying the data format and attributes, multiple data groups are handled virtually as data for one person. Then, this data group is assigned a token, which enables it to be acquired as a batch.

As a result, data for multiple people can be acquired as a batch in two sequences irrespective of the number of people's data to be acquired, while taking advantage of the access control safety inherent in the UMA protocol. In UMA 2.0, because consent is confirmed on the authorization server, only the data with consent is transferred even when data is acquired as a batch between organizations. The history of an individual's consent is difficult to falsify because it is stored as a hash value in the blockchain.

However, this extension also creates vulnerability to attacks where fraudulent links are sent via targeted emails to acquire data. As a measure against such attacks, a token verification function has also been added.

4. Utilization examples

This section describes two utilization examples related to Chain Data Lineage.

4.1 Telematics automobile insurance

The current trend in the insurance industry is for insurance premiums to be calculated very precisely according to risks. Meanwhile, the advance of ICT and sensor technology enables detailed driving habits data to be collected. In telematics automobile insurance, the driving habits data of the automobile is collected and analyzed to calculate an insurance premium suitable for the individual driver, allowing the provision of insurance optimized for the risk. This creates benefits for both the insurance company and the driver. To achieve further optimizations in the future, the insurance companies will collect additional information such as driving habits

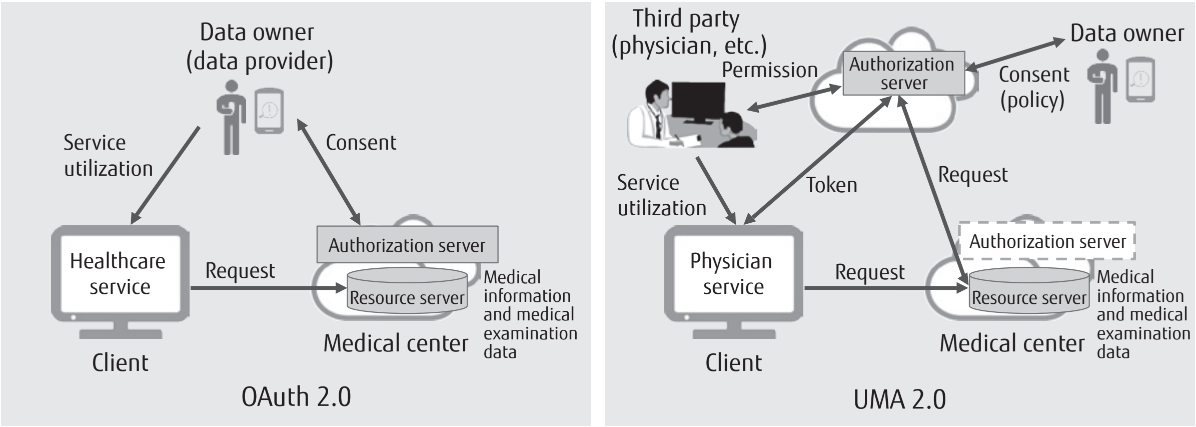

Figure 4
Examples of OAuth and UMA configurations.

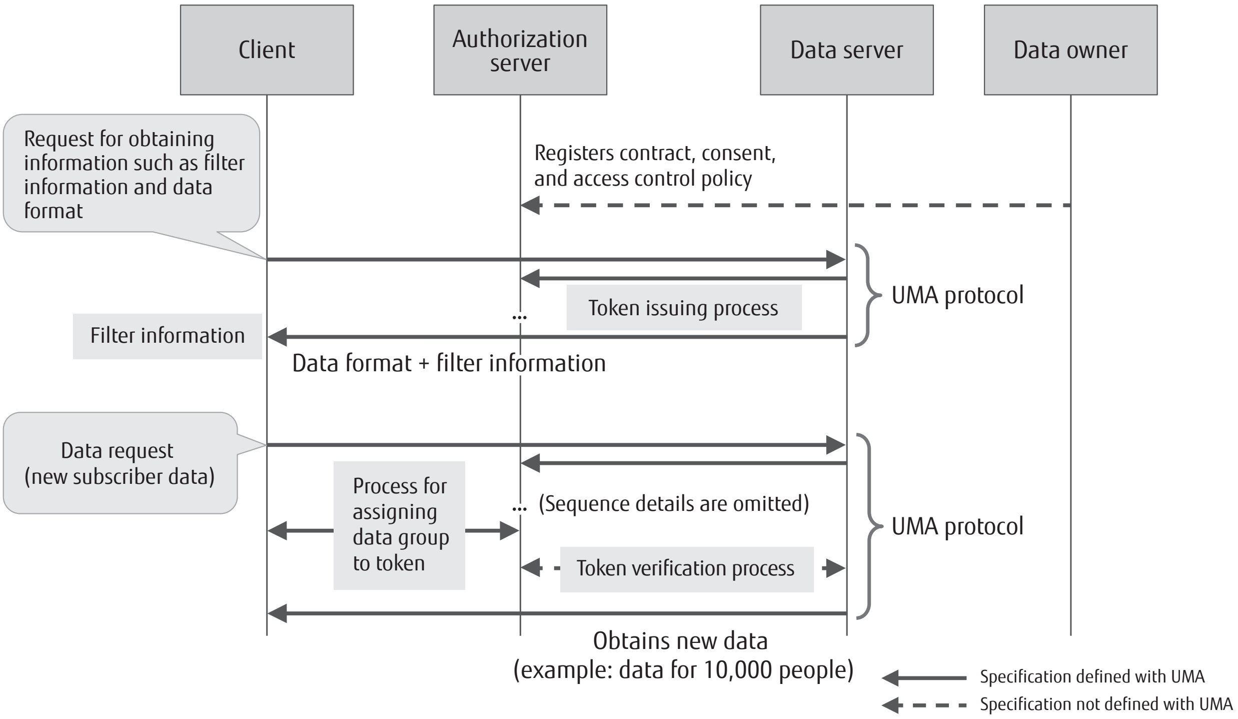
Client
Authorization server
Data server
Data owner
Request for obtaining information such as filter information and data format
Registers contract, consent, and access control policy
Token issuing process
UMA protocol
Filter information
Data format + filter information
Data request (new subscriber data)
Process for assigning data group to token
(Sequence details are omitted)
UMA protocol
Token verification process
Obtains new data (example: data for 10,000 people)
Specification defined with UMA
Specification not defined with UMA

Figure 5
Processing sequence with UMA 2.0 extension.

data even when driving a car other than one's own and driving conditions like weather. By combining these, the more advanced analysis seems to become possible.

In this way, the insurance industry will start using data in novel ways, as well as using processed data from new data analysis vendors that will offer advanced analysis. In Chain Data Lineage, the data history information can be checked by the data users, which enables them to determine in advance whether the data is reliable.

Figure 6 shows an automobile insurance system based on driving habits data. In this system, the information up to calculating the driver model data is disclosed to the data user at the insurance company as history information. This history information contains data such as the automobile manufacturer from which the driving habits data was provided, the analysis vendor that analyzed the data, and how the driver model data was calculated. Previously, the insurance company could only check the information in the driver model data of the previous step. But by checking this history information, the insurance company can now understand the background to the data, which greatly increases data reliability.

4.2 Utilization at pharmaceutical companies

A wide range of case studies of medical examinations must be collected and analyzed when manufacturing pharmaceutical products. Similarly to the example of the telematics automobile insurance in the previous section, Chain Data Lineage can be applied to the process for exchanging data between the various stakeholders of the pharmaceutical business. Medical examination data generated at hospitals and other institutions is analyzed at universities and data analysis vendors, and the results are used by pharmaceutical companies. Pharmaceutical companies can use Chain Data Lineage to check history information as soon as the data is generated, which enables the reliability of the data to be confirmed before using it in drug manufacture. Proper handling of medical examination data is important for the privacy of patients; therefore the consent of patients must be obtained before transferring any such data. The consent management technology of Chain Data Lineage can be applied to this process as well.

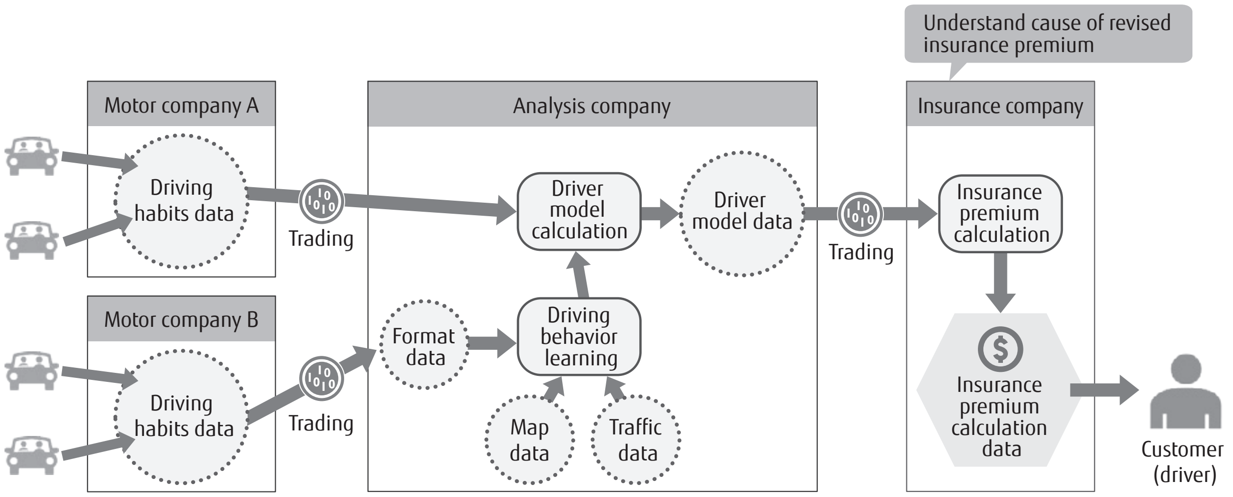

Figure 6
Automobile insurance system based on driving habits.

5. Conclusion

This paper described the Chain Data Lineage technologies used to implement data history management technology and individual consent management in order to ensure reliability of data generated and processed across multiple organizations. These technologies enable people to use large amounts of data with peace of mind, which builds a data society on the basis of trust.

In the future, Fujitsu will increase the application of Chain Data Lineage according to specific services that involve data utilization and analysis. Fujitsu will also promote the standardization of data utilization in order to further disseminate these technologies.

The objective is to apply these technologies to greater amounts of data so that specialists at a large number of companies can analyze and process the data, creating an environment where people can use data with peace of mind.

All company and product names mentioned herein are trademarks or registered trademarks of their respective owners.

References

1) Ministry of Internal Affairs and Communications: “Information and Communications in Japan (White Paper 2012).” (in Japanese). *http://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h24/*

2) M. Matsubara et al.: Data Management System that Facilitates the Value Creation Cycle. FUJITSU Sci. Tech. J., Vol. 54, No. 5, pp. 30–36 (2018). *https://www.fujitsu.com/global/documents/about/resources/publications/fstj/archives/vol54-5/paper05.pdf*

3) Fujitsu: Fujitsu Develops Technology to Improve Reliability of Data Distribution across Industries. *https://www.fujitsu.com/global/about/resources/news/press-releases/2018/0920-02.html*

4) Fujitsu: Fujitsu Develops Blockchain-based Software for a Secure Data Exchange Network. *https://www.fujitsu.com/global/about/resources/news/press-releases/2017/0605-01.html*

5) Y. Ejiri et al.: Data Distribution and Utilization Services Across Business Boundaries. FUJITSU Sci. Tech. J., Vol. 55, No. 3, pp. 9–14 (2019). *https://www.fujitsu.com/global/documents/about/resources/publications/fstj/archives/vol55-3/paper02.pdf*

6) Fujitsu: Virtuora DX Data Distribution and Utilization Service (in Japanese). *https://www.fujitsu.com/jp/products/network/carrier-router/dataexchange/virtuora-dx/saas/*

7) Internet Engineering Task Force (IETF): RFC 6750 “The OAuth 2.0 Authorization Framework: Bearer Token Usage.” *https://tools.ietf.org/html/rfc6750*

8) Kantara Initiative: User-Managed Access (UMA) 2.0 Grant for OAuth 2.0 Authorization. *https://docs.kantarainitiative.org/uma/wg/rec-oauth-uma-grant-2.0.html*

Masazumi Matsubara
Fujitsu Laboratories Ltd.
Dr. Matsubara is currently engaged in the research and development of basic technologies for data distribution and utilization.

Takeshi Miyamae
Fujitsu Laboratories Ltd.
Mr. Miyamae is currently engaged in the research and development of basic technologies for blockchain and data distribution and utilization.

Akira Ito
Fujitsu Laboratories Ltd.
Mr. Ito is currently engaged in the research and development of data distribution and utilization.

Ken Kamakura
Fujitsu Laboratories Ltd.
Mr. Kamakura is currently engaged in the research and development of data security and blockchain applications.