Data Management System that Facilitates the Value Creation Cycle

● Masazumi Matsubara ● Minoru Nakamura ● Mitsuru Sato ● Eiji Yoshida
● Naoki Matsuoka

Data distribution and utilization are being consolidated due to support by legislation such as the Basic Act on the Advancement of Public and Private Sector Data Utilization and the Amended Act on the Protection of Personal Information and also by the establishment of the Data Trading Alliance consisting of industrial, governmental, and academic institutions. As corporate businesses become increasingly digitalized and connectable with potential partners in the future, Digital Co-creation among different industries will accelerate more than ever and give rise to innovations. Fujitsu Laboratories propose a concept, Connected Digital Place, which enables Digital Co-creation among various industries. As a core part of the concept, we have been conducting research and development of a data-driven platform that manages data, converting its formats into connectable ones. This paper introduces the latest research to realize the data-driven platform, and Fujitsu's approach to promoting the value creation cycle through the use of this platform.

1. Introduction

The movement to promote innovations by utilizing data of multiple companies has been gaining momentum. In Japan, a broad range of legislation has been put in place from late 2016 to 2017: the Basic Act on the Advancement of Public and Private Sector Data Utilization, which promotes the distribution and utilization of data owned by public and private organizations, and the Amended Act on the Protection of Personal Information, which enforces the secure use of personal information. Furthermore, the Data Trading Alliance consisting of industrial, governmental, and academic institutions was founded in November 2017. Several companies such as General Electric and Komatsu quickly found value in data utilization and have been leading business in this area by providing their own data utilization platforms and building ecosystems on these platforms.

We, at Fujitsu Laboratories, are also actively conducting research and development in this area, proposing a concept, Connected Digital Place, which realizes Digital Co-creation among companies by connecting businesses of different industries. Connected Digital Place is a space for the creation of brand new businesses by digitizing and connecting businesses of customers in various fields such as government, education, manufacturing, and healthcare. Likewise, Germany's Industrie 4.0, the Industrial Internet proposed by the Industrial Internet Consortium and Japan's Society 5.0 aim to realize further economic growth and a society where people can live comfortably and prosperously by connecting the systems of various industries.

In any of the above-mentioned system linkages including Connected Digital Place, Digital Co-creation among companies belonging to different industries is realized by data connection. However, because the structure and semantics of data are not unified among companies, connecting data of different companies is difficult even if the data are of the same type. Thus, how to quickly provide data formatted as each company wants is a common urgent problem for all data utilization platforms. We are tackling the problem by developing a data-driven platform as a main component of Connected Digital Place.

In this paper, we first outline the data-driven platform and then explain its components: the Data Bazaar and the data infrastructure. Next, we introduce the

activities undertaken toward the practical application of this platform to the value creation cycle.

2. Data-driven platform

Figure 1 shows the overview of the data-driven platform. The key factor of Connected Digital Place is the diversity of data whose handling varies depending on the systems of customers. The data-driven platform combines, analyzes, and refines data into valuable information for businesses in accordance with business objectives and contracts. Then, new businesses and service are created as digital innovation, and this promotes economic and industrial growth.

The platform consists of Data Bazaar and the data infrastructure. Data Bazaar supports data management, distribution, and use. Data gathered from real-world entities such as people, cars, and factories are stored in the data infrastructure. Then, after processing and distribution by Data Bazaar, data from different industries are combined and used according to business proposes, such as creating new digital businesses, improving existing digital businesses, and so on.

As feedback to the real world, new digital businesses built in the virtual world yield new value in various ways such as navigating human activities, sharing vacant warehouses, and improving factory utilization. Meanwhile, new data such as data access logs and business transaction logs are generated through the operation of digital businesses. The platform manages, processes, and analyzes such data appropriately, thereby improving existing businesses and utilization of such data for new businesses.

In order to realize the above, data must be converted into a form that permits connection. For example, when the data of a company is described according to that company's proprietary product ID system, it must be converted into data that conforms to the industry standard product ID system. At Fujitsu Laboratories, we call this "connectable information."

To distribute and utilize data as connectable information, it is necessary to process the data so that other companies can use them and extract metadata so that they can be searched. It is also essential to enhance the data processing infrastructure—this includes things such as managing the quality and quantity of data, large-scale data processing, and real-time data processing. Moreover, a high-speed and large-capacity data infrastructure is required as the infrastructure supporting such data distribution and utilization.

The following sections introduce the latest technologies developed for Data Bazaar to support data distribution and utilization, and the underlying data infrastructure.

3. Data Bazaar

Data Bazaar is a collective term for the technologies on the data distribution and utilization platform developed by Fujitsu Laboratories (**Figure 2**). Data Bazaar facilitates data utilization by three major automated features: participation in the field of data distribution and utilization, easy and secure data

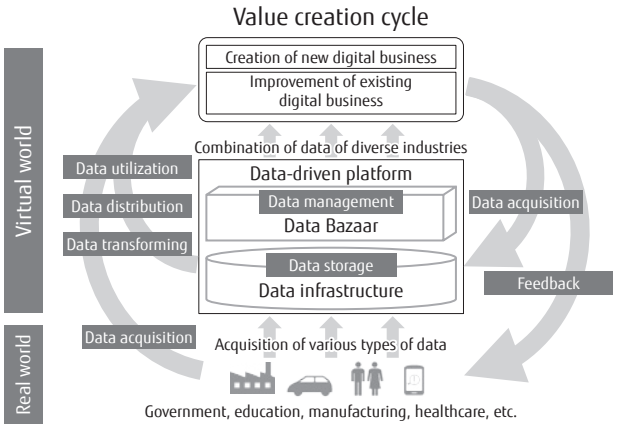
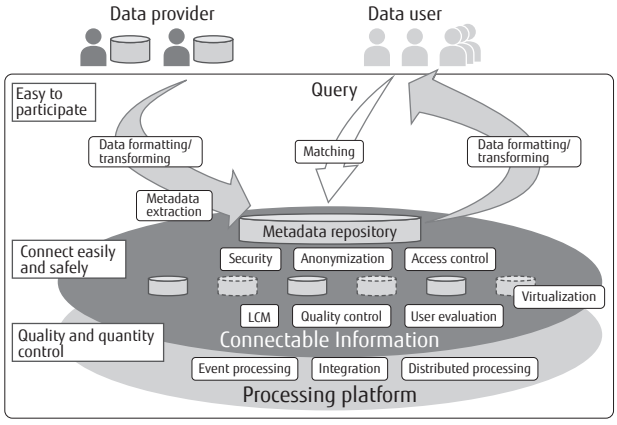


Figure 1
Data-driven platform.



LCM: Life cycle management

Figure 2
Data Bazaar.

linkage, and data quality and quantity management. Formatting data into connectable information and distributing it are important for these features.

For the various data such as corporate data and open data released by national governments and local governments to be available as connectable information to companies, satisfying the following two points is particularly important.

1) From among a large amount of data, being able to find data that can be combined with one's company's data to create value.
2) Being able to convert data to one of the data formats and unify the semantics of the data to allow data combination.

However, data combination trials, data format conversion, and unification of semantics are extremely time-consuming operations that require skilled engineers. This is a bottleneck that hinders data utilization.

Fujitsu Laboratories' Data Bazaar aims to support not only skilled engineers by reducing required work time but also unskilled users by automating typical analysis tasks. The ultimate goal of Data Bazaar is to automate series of data utilization related processes and to offer customers concierge-style services that allow them to effectively combine and utilize data.

In the following, as technologies for realizing automation and high speed execution of such processing, we introduce automatic metadata extraction and search technology, technology to automate data prepping, and technology to speed up data analysis processing.

3.1 Automatic metadata extraction and search technology

Turning the data described in point 1) above into searchable and connectable information requires recording human knowledge, such as knowledge relating to structure, characteristics, and creation purpose, as metadata that programs can read. Such metadata include data types, meanings of data strings, tags, and statistical feature values. Storing these metadata in a repository makes searches and recommendations possible. For example, interactive search technology being developed at Fujitsu Laboratories analyzes sentences in metadata using natural language processing technology, and recommends appropriate categories to users. As a result, it is possible to quickly acquire relevant data from a huge data set even with ambiguous search words.

That said, the creation of metadata is extremely costly both time-wise and money-wise. Thus, the automation of metadata generation is essential to handle large amounts of data. To solve this problem, Fujitsu Laboratories is developing technology to automatically generate metadata based on similarities with registered data and the records of data processing flow, and so on.

3.2 Data prepping automation technology

The data of companies come in various file formats such as CSV, JSON, and XML, and a variety of data formats are in use. Data such as persons' names, company names, and addresses themselves differ in terms of system, and data notations are not necessarily consistent. To convert the data mentioned in point 2) above into connectable information that can be combined, data prepping to make the data unified in terms of format and semantics is required. However, such work is often done manually and accounts for 80% of data analysis processing, which is a serious obstacle to advancing data utilization.

The problem with data prepping is that it takes time. On the other hand, engineers can envision the data after processing even if the processing procedure is unknown. Fujitsu Laboratories has developed Programming by Example (PBE) technology that performs reverse synthesis of processing procedures through AI technology that performs graph searches and proprietary pruning, using several samples of processed data written by engineers.[1)] Application of this technology to prep data from about 8,000 past points of sale (POS) transactions, a task that until now would take five days, shortened the processing time to about half a day.

3.3 Technology to speed up data analysis processing

Even when data analysis work can be automated, final judgment by humans is required. However, the repeated trial and error approach that produces the best results will be avoided if it takes one whole night to execute a single data analysis process. Therefore, it is necessary to realize real-time analysis by speeding up data analysis processing as much as possible.

Fujitsu Laboratories has been studying ways to speed up analysis using query languages such as SQL, which are often used in data analysis processing. Until now, we have been developing technology to speed up database components–for example, by developing a parallel query execution function and a column-oriented storage function for PostgreSQL, and by offloading query processing for PostgreSQL, MongoDB, and Shunsaku[note)] to Apache Spark, a distributed execution environment.

Fujitsu Laboratories focuses on cloud-scale distributed execution technology as a next-generation data processing acceleration method. Until now in conventional on-premise systems, about three to ten cluster systems were introduced for each customer. However, on multi-tenant distributed execution platforms, multiple customers share a computing resource pool of one hundred to one thousand units. Since the timing of data processing differs according to each customer, data processing is sped up by allowing customers to use distributed execution technology that gives them exclusive access to all the resources that are available within the pool at any given point in time. With such technology, scheduling and the avoidance of conflicts among multiple customer queries are important. Fujitsu Laboratories is currently studying a multi-tenant distributed execution platform based on Spark.

4. Data infrastructure

In a data-driven platform, an infrastructure for storing data acquired from people, cars, factories and the like in the real world and making this data available to Data Bazaar for processing aimed at distribution is essential. An infrastructure supporting such data distribution requires functions for the efficient storing of large amounts of data and functions to process data at high speed. Fujitsu Laboratories calls the infrastructure that has the above functions data infrastructure, and is currently developing the technologies required for realizing this data infrastructure.

Figure 3 shows the structure of the data infrastructure. To efficiently store and provide various types of data distributed in different environments, the data infrastructure must be flexibly configurable according to the location and application, and it must allow various types of data to be linked. Further, to realize faster data analysis, the data infrastructure must have dramatically high data access performance as compared with traditional storage systems. To meet these requirements, the data infrastructure must offer a structure in which each layer is seamlessly integrated:

- ultra-high-speed data stores capable of handling hot data manipulated with high frequency,
- high-speed storage for warm data that is not manipulated so frequently but is voluminous, and
- large-capacity storage capable of efficiently handling large amounts of cold data manipulated with low frequency.

The data is to be migrated to the required locations and layers as necessary and to be supplied at the appropriate time. Data migration is to be automatically performed according to the operation of the workload that uses the data.

In order to flexibly construct the data infrastructure and realize free data migration between layers, the entire data infrastructure is composed of software-defined storage (SDS). The use of SDS makes it possible to provide data stores integrated with a processing engine–for example, an ultra-high-speed data store composed of dynamic random access memory (DRAM) or non-volatile memory (NVM). Further, a wide area distributed analysis engine for analyzing workload's

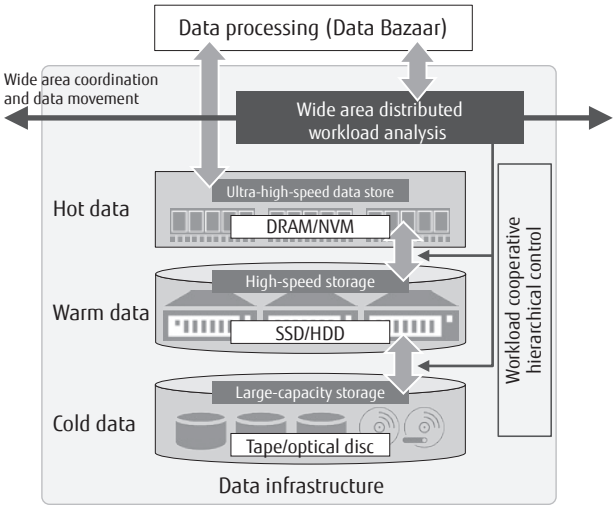


Figure 3
Data infrastructure.

note) FUJITSU Software Interstage Shunsaku Data Manager, an XML database engine.

behavior in real time and enabling seamless data access is provided. As a result, data can be automatically migrated to different layers and locations according to the operation of the workload.

Fujitsu Laboratories is developing component technologies to realize the data infrastructure. In the following, we introduce large-capacity memory technology for realizing ultra-high-speed data stores and workload analysis technology for realizing automatic data migration between layers.

4.1 Large-capacity memory technology

As CPUs become increasingly fast and data analysis capabilities improve, faster data input and output speed for analysis is also required. In-memory processing in particular is a method used when especially high-speed processing is required. Placing all the data in the memory reduces the input/output overhead of the storage and achieves high-speed processing. However, conventional in-memory processing has the problem that the data size that can be handled is small because the amount of data that can be analyzed is limited by the memory capacity of the server.

To solve this problem, Fujitsu Laboratories is studying a method to expand the data size that can be handled by in-memory processing by treating next-generation NVM as large-capacity main memory.[2] This method makes it possible to increase the amount of data that can be handled by in-memory processing by a factor of 10 without substantial performance penalty, allowing larger amounts of data to be processed at high speed.

4.2 Workload analysis technology

Seamless connection from a memory-based ultra-high-speed data store to a large-capacity data store requires arrangement of the data according to the usage status of the workload. To achieve this, it is necessary to first analyze data migration on the data infrastructure side and control it according to how the data is actually used. Fujitsu Laboratories also conducts research and development on data access analysis technology for workloads. This technology can be applied also to existing systems–for example, it can be used for analysis of storage bottlenecks for virtual desktop.[3]

In the data infrastructure, we apply this technology to analyze how data is used, and automatically arrange the necessary data in the optimum data layer. The aim in doing so is to realize a system that can store large-capacity data at low cost and can provide it to analysis engines at high speed.

5. Implementation example of value creation cycle

In parallel with the development of the data-driven platform mentioned in the previous section, we are conducting demonstration experiments for business verification with customers. This section introduces as an example a demonstration experiment in the regional government field between Shimane Prefecture, Payke, Inc. (hereafter, Payke), and Fujitsu through Digital Co-creation (**Figure 4**).

One of the challenges of regional government is regional revitalization. By connecting the data of the government and enterprises, it is possible to gain insights into the behavior of local people and things that could not be visualized so far. In turn, by utilizing the results for the businesses of local companies, it is possible to revitalize local industries. To demonstrate this, Fujitsu Laboratories used its data-driven platform to create an environment for the combined utilization of Shimane Prefecture's prefectural product authentication data such as "Oishimane Certification"[note1] and "Shimane Furusato-Shokuhin Certification"[note2] data, and Payke's own data.

Using this data-driven platform, since September 2017, the Shimane Prefectural Products and Tourist Center in the city of Matsue in Shimane Prefecture and Tokyo's Shimane-Kan regional specialty shop in Nihonbashi have been offering services introducing products of the prefecture to foreign tourists. Foreign tourists can check a given product's information, its prefectural product certification mark, and "special feature" type information from the producer in their mother tongue by holding a camera-equipped in-store

note1) A certification program for safe and delicious agricultural, forestry, and fishery products by Shimane Prefecture's Governor. *Oishimane* is a coined term consisting of "*oishii* (the Japanese word for delicious)" and "*Shimane*."

note2) A certification program for genuine food products of Shimane Prefecture.

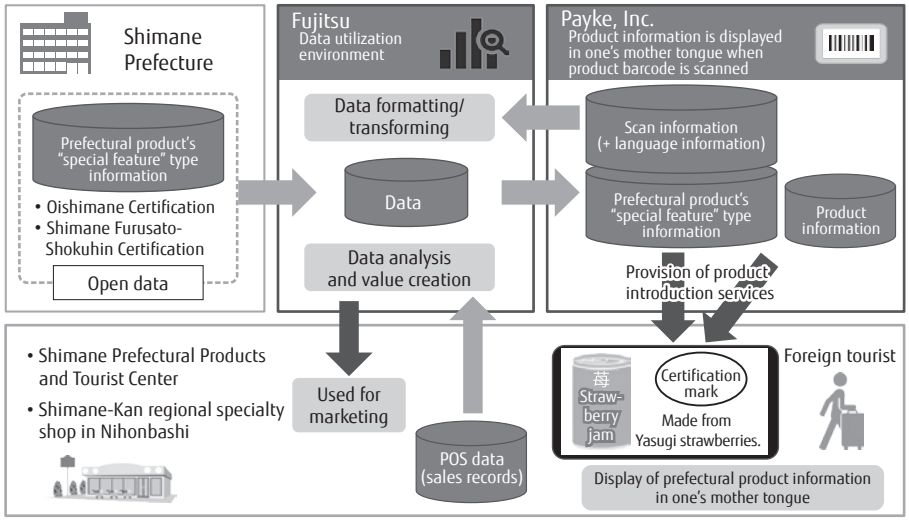


Figure 4
Demonstration experiment in field of regional government.

terminal over the barcode attached to that in-store item. This allows information on prefectural products to be conveyed to foreign tourists in more detail, improving recognition of Shimane Prefecture and its products, promoting the purchase of prefectural products, and revitalizing local industries.

On the other hand, the data-driven platform makes it possible to grasp what kinds of people in which countries are interested in which prefectural products. Furthermore, this platform associates the in-store product lookup history, prefectural product certification data, and sales data at the Shimane Prefectural Products and Tourist Center and Shimane Prefecture's antenna shop in Tokyo. This enables fine analysis of customer behavior such as identification of items customers were not interested in at first and things that they were interested in but did not purchase. This in turn has made it possible to design marketing measures more closely tailored to customers, creating a virtuous cycle of value creation.

Looking ahead, to revitalize regional business as a whole, we are also studying the combination of data relating to tourists' movements and sales data of other retailers, and providing data to companies in different industries.

6. Conclusion

In this paper, we introduced Fujitsu Laboratories' activities to develop a data-driven platform that is the core of Connected Digital Place for realizing Digital Co-creation among companies in different industries. We also introduced an actual case of creating a virtuous cycle of value creation.

Going forward, we will continue to develop technologies for realizing Digital Co-creation centering on data, and by providing demonstration experiments in cooperation with customers, we will quickly provide technology that can truly support our clients' businesses.

References

1) Fujitsu Laboratories: Fujitsu Automation Technology Preps Data to Accelerate Analysis.
http://www.fujitsu.com/global/about/resources/news/press-releases/2017/0915-01.html
2) Fujitsu Laboratories: Fujitsu Develops Optimized Software-Controlled Solid-State Drive for Big Data Processing.
http://www.fujitsu.com/global/about/resources/news/press-releases/2015/1119-01.html
3) Fujitsu Laboratories: Fujitsu Develops Automated Analysis Technology to Identify Causes of Performance Degradation in Virtual Desktops.
http://www.fujitsu.com/global/about/resources/news/press-releases/2017/0523-01.html

Masazumi Matsubara
Fujitsu Laboratories Ltd.
Dr. Matsubara is currently engaged in research and development of digital service architecture.

Minoru Nakamura
Fujitsu Ltd.
Mr. Nakamura is currently engaged in research and development of databases.

Mitsuru Sato
Fujitsu Laboratories Ltd.
Dr. Sato is currently engaged in research and development of storage software and storage server architecture.

Eiji Yoshida
Fujitsu Laboratories Ltd.
Dr. Yoshida is currently engaged in research and development of data system technology.

Naoki Matsuoka
Fujitsu Laboratories Ltd.
Mr. Matsuoka is currently engaged in research and development of digital business creation.