KaaS Knowledge Service Platform Facilitating Innovation in Social Infrastructure

Ryoichi Ueda
Yoshinori Sato
Masakatsu Mori
Kozo Nakamura
Nobutoshi Sagawa

OVERVIEW: Hitachi is working on research and development aimed at implementing KaaS that extract and present valuable information (knowledge) from the large volumes of data generated by social infrastructure such as industry, transport, and electricity supply. Two issues associated with realizing this goal are getting access to extensive computing capacity and ensuring the privacy of the data. To solve these problems, Hitachi is developing an architecture that can run various different analysis algorithms on large distributed processing platforms and, in addition to the ability to increase capacity simply by adding computing resources, has also developed privacy protection systems that incorporate two features: data access control for multiple items with multiple IDs and k-anonymization. The resulting technology is currently being used for applications such as maintenance and diagnostic services and location-aware services where it is the subject of ongoing evaluation and improvement. The aim for the future is to provide knowledge platforms that merge information in a way that transcends boundaries between industries by building systems with a high level of computing performance able to analyze data from across different industries.

INTRODUCTION

WHILE most of the things that managers have sought to achieve with IT (information technology) in the past have related to business efficiency improvements achieved through the replacement of human labor, in more recent years these expectations have been undergoing a steady shift from business efficiency improvement toward the creation of high added value.

When one considers the changes in IT itself, examples include the widespread adoption of advances in sensor technologies such as RFID (radio-frequency identification) and GPS (global positioning system), improvements in communication technologies that can collect the huge volumes of data produced by these sensors, the establishment of broadband communication infrastructure, and advances in virtualization technologies and cloud computing that provide extensive computing resources at affordable cost.

Against this background, Hitachi is proposing to provide KaaS (knowledge as a service) and has already embarked on research into platform technologies. Such services can extract and present information (knowledge) with high added value by analyzing the huge volumes of raw data available from sources such as GPS and other sensors or the log information recorded by IT systems. This article describes a KaaS knowledge service platform that contributes to

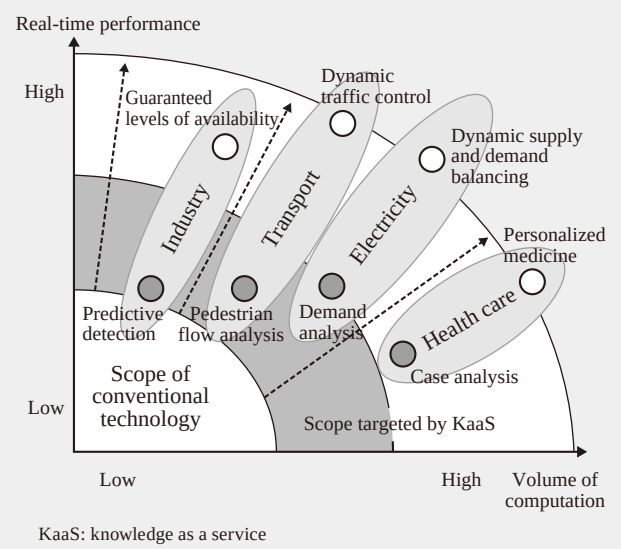

Fig. 1—Characteristics of Example KaaS Applications in Industry, Transport, Electricity, and Health Care.
The aim for KaaS is its use for applications that are difficult to implement using conventional technology because of the volume of computation and real-time performance requirements.

innovation in social infrastructure.

POTENTIAL OF KAAS

Rather than improving the business of specific companies, Hitachi's objective for KaaS is to add significant value to social infrastructure. Put another way, it is seeking to deliver innovation. In the industry field, for example, the use of advanced algorithms to analyze data from sensors attached to equipment can predict equipment malfunctions before they occur and enable business models that rely on a guaranteed level of equipment availability. In the transport field, the analysis of data from sources such as electronic ticketing, probe cars, and mobile GPS could be used not only to determine the macro-movement of groups of people but also to clarify and predict the micro-movement of individuals. The availability of such information could then provide the basis for traffic control or other innovative services that have not existed before. Other potential applications include establishing dynamic supply and demand balancing plans designed to minimize environmental impact and ensure a reliable electricity supply over a wide area based on the integrated analysis of data from different industries such as the level of electricity generation at electricity utilities or within the community and predictions of electricity consumption derived from the production and sales plans of manufacturing and distribution businesses.

For KaaS to realize these desired innovations in social infrastructure, it will be necessary to analyze efficiently the large volumes of data, including personal data that is collected by operating companies and other service providers with an important public role.

To satisfy this requirement, Hitachi is focusing on two major technical issues in particular. The first issue is how to extract valuable information (knowledge) efficiently from the large volumes of data originating from social infrastructure systems and it is related to the IT architecture. The second is the issue of privacy and security and how to make use of data that may contain personal or confidential information without putting it at risk.

Extraction of Knowledge from Large Volumes of Data

In addition to the large volumes of data that need to be processed, another important consideration, if valuable information (knowledge) is to be extracted from the data accumulated by companies that support social infrastructure, is the need to use techniques such as machine learning that impose a heavy processing load. Also, because a high level of real-time performance is required if the extracted knowledge is to be used in the control of equipment, how to provide the extensive computational resources that are required and how to utilize those resources efficiently are also issues. Fig. 1 shows typical examples of KaaS applications in different fields and their characteristics.

The aim for KaaS is to implement applications that are difficult to achieve using conventional computing technology because it would be too expensive or because of its inability to provide sufficient accuracy given the processing volume and real-time performance requirements.

Privacy Protection

Terms like "lifelog" and "personal information" are used to refer to real-world data about the lifestyles of individual people and much attention has been focused on issues such as potential uses for this information and how to keep it secure[(1)]. Although it is considered that more time will be needed to determine how best to handle lifelog and similar data and to establish a public consensus on the issue, at a minimum, managing the data on the principle of protecting personal information will be essential.

On the other hand, approximately 70% of information security incidents are caused by errors such as data being misplaced or sent to the wrong address and cases of misuse by authorized users of the data have included the release of more than one million items of personal information at a time[(2)]. Given the nature of these incidents, the objective is to prevent potential risks such as information leaks or unauthorized use before they occur, starting with technical measures. Technologies for achieving this include authentication, cryptography, and access control.

Further, to deal with the case when an incident does occur, it is also desirable to have procedures in place beforehand to minimize the consequences. In other words, the life cycle of data such as lifelog extends from collection through to transmission, knowledge extraction, knowledge utilization, and disposal, and it is necessary to establish new risk prevention measures at each of these stages that will keep the data as safe as possible. However, measures for keeping the data safe need to be designed in a way that does not obstruct its use in processing such as knowledge extraction.

APPROACH TO SOLVING PROBLEMS

Processing Techniques for Extracting Knowledge from High Volumes of Data

Services like KaaS that are provided over the web are typically configured with a three-tier web-based system structure consisting of web, AP (application), and DB (database) servers. However, the three-tier web-based structure is suitable for systems with comparatively small data volumes and simple computation and is not suitable for KaaS systems that handle large volumes of data and perform complex computations such as machine learning functions.

Also, KaaS systems are designed to work with a wide range of different data derived from RFID, GPS, IC (integrated circuit) cards, and other such sources and representing the location or other details of real-world people and objects, and they need to be capable of performing analyses from a range of different perspectives.

Hitachi has adopted a configuration in which the DB server tier of the conventional three-tier web-based structure is replaced by the KaaS-Base data management tier based on the MapReduce distributed processing technique to satisfy the "data volume" requirements, and in which the AP server tier is replaced by the KaaS-Core knowledge processing tier based on a framework for the integrated and efficient handling of various different types of data to satisfy the "data type" requirements (see Fig. 2).

The MapReduce framework used in KaaS-Base is a parallel distributed processing technique devised by Google* in which programs are written in pairs of "map" steps and "reduce" steps. Parallel processing is made possible by writing the program in such a way that each map step is independent of other map steps and by splitting up the input data.

The adoption of the MapReduce distributed processing technique provides a system configuration with a high level of scalability that can cope with future increases in data volume simply by adding computing resources and without needing to make program changes. The intention for the KaaS system is to use the Hadoop open-source implementation of MapReduce as a base and optimize it to ensure adequate scalability for practical applications. KaaS-Base also incorporates a stream processing engine to allow the use of real-time models.

For KaaS-Core meanwhile, a set of components for analysis algorithms and similar operations and a library organized by data type of functions shared between application developers is being prepared based on a study of the type and usage patterns of actual data.

The aim is to make it easy for application developers to create business applications that extract valuable information from large volumes of data by combining the analysis framework, analysis components, and other resources in the KaaS system.

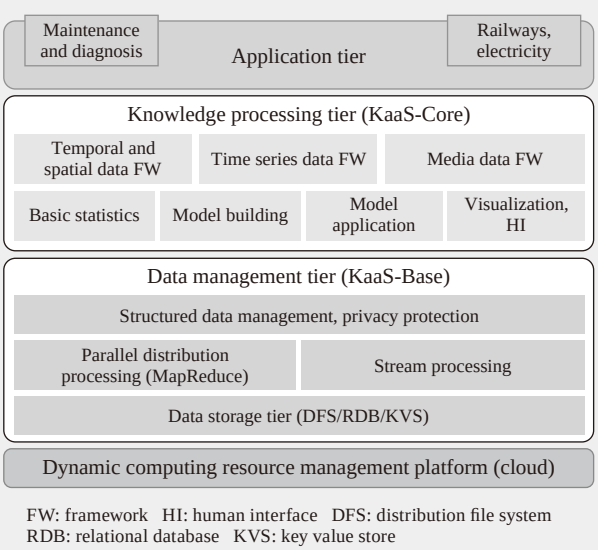

Fig. 2—KaaS System Architecture.
The architecture is made up of the KaaS-Base data management tier based on the MapReduce distributed processing technique and the KaaS-Core knowledge processing tier based on a framework for the integrated and efficient handling of various different types of data.

Privacy Protection Platform

One effective approach to resolving the issues surrounding privacy protection is the use of "anonymizing" techniques that make data safe by making it impossible to identify individuals' identities. Fig. 3 shows an overview of the anonymizer platform corresponding to "privacy protection" in Fig. 2.

The anonymizer platform manages lifelog and similar data by dividing it into "base items" consisting of name, address, age, gender, and other items of personal information that on their own or in combination can identify individuals and "sensitive items" which include everything else. The main functions implemented by the platform are (1) item-specific ID (identifier) control and (2) information granularity control[3].

The item-specific ID control function restricts simultaneous access to fields that cover both base items and sensitive items and assigns multiple IDs to records. Although implementations vary depending on the type

* Google is a trademark of Google Inc.

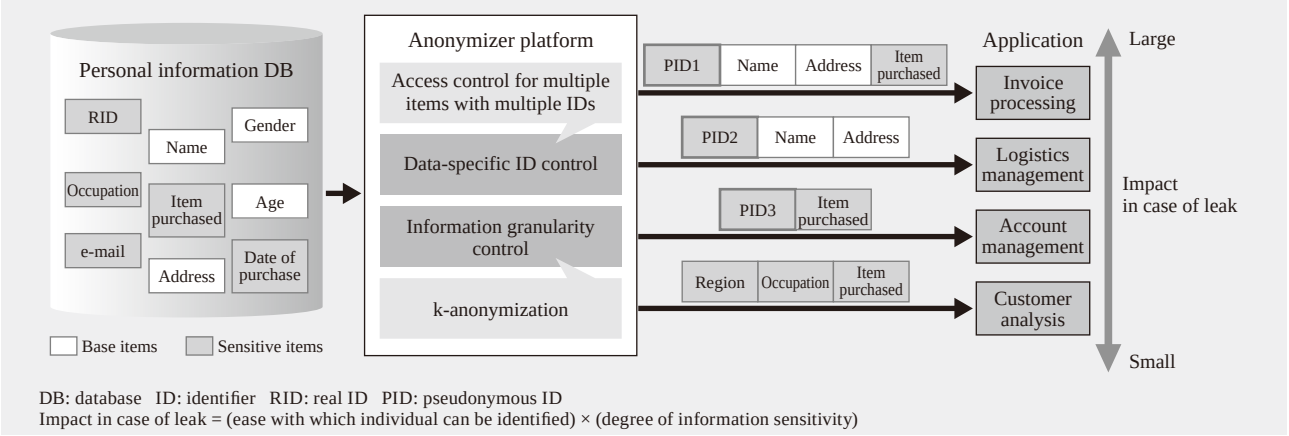

Fig. 3—Overview of Privacy Protection Platform.
The information used can be minimized and unauthorized linking of data prevented by controlling the personal ID, available data items, and granularity of data item values independently for each application.

of system, the essential element is the inclusion both of simultaneous access control and use of multiple IDs based on the concept of "privacy by design."

It is anticipated that potential KaaS applications will want to have safe access to specific subsets of the base items such as for the analysis of particular combinations of gender, partial address, and location information records. It is in cases like this that the information granularity control function becomes necessary and the way this is implemented is called "k-anonymization." k-anonymization is a technology for controlling data granularity (vagueness) so that there are at least k instances of a particular item in any data set. Hitachi is working on the development of k-anonymization techniques suitable for use in realistic scenarios such as providing a way to assign a priority to items that need to be subject to granularity control.

In this way, anonymization plays a particular role as a way of protecting the privacy of data such as lifelog. However, it is important to remember that technology cannot resolve all of the issues. It is also essential that the system is operated in a way that complies with the demands of society by, for example, obtaining necessary permissions to use data when it is collected and only using data in accordance with the stated purpose of use.

EXAMPLE KAAS APPLICATIONS

This section describes an example KaaS application that provides diagnostic services in the field of maintenance.

The maintenance diagnostic service involves fitting temperature, pressure, and other sensors to the equipment to be maintained and analyzing the status data collected by these sensors to detect outliers in the equipment. The aim is to reduce maintenance costs and avoid downtime caused by unexpected faults.

The analysis of the sensor data uses machine learning technology to perform two steps: (1) determine the range of data values that constitutes normal operation based on sets of sensor data collected when the equipment is functioning correctly, and (2) calculate the extent to which the current sensor data has diverged from this range. The system reduces downtime due to unexpected faults by deeming the equipment to be "abnormal" if this divergence exceeds a predefined threshold based on the equipment characteristics. This in turn triggers maintenance work such as replacing parts.

Hitachi is in the process of implementing this maintenance diagnostic service application and the required analysis algorithms on a KaaS platform to configure a system able to analyze large volumes of sensor data at high speed (see Fig. 4).

For the future, Hitachi believes it can contribute to improving the efficiency of a wide range of maintenance work including fault cause analysis and even the provision of replacement parts by collecting large amounts of abnormal data together with information about the underlying cause of the outlier and how to resolve it.

In addition to KaaS applications such as this one that use time series data, other services that Hitachi is investigating include analysis services for text, audio, and other multimedia data, location-aware services, and real-time services that provide information about the movement of people in specific areas based on temporal and spatial data such as GPS and electronic ticket logs.

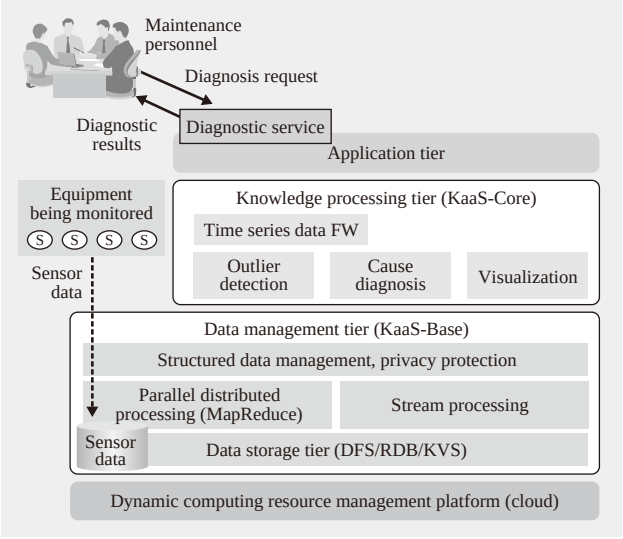

Fig. 4—Overview of Maintenance Diagnostic Service.
The service collects data on the status of the equipment being monitored from sensors fitted to the equipment and analyzes it to detect outliers and identify the cause of the faults that occur.

CONCLUSIONS

This article has described a KaaS platform that contributes to innovation in social infrastructure.

Hitachi started applying its KaaS system to a range of different real-world data during the 2010 financial year. Hitachi intends to make improvements based on the knowledge it gathers from these activities and continue working toward commercial applications.

In the future, Hitachi aims to supply a knowledge platform that can integrate knowledge in a way that transcends the boundaries between industries to conduct analyses that combine data from different industries such as linking the production and sales plans of manufacturing and distribution businesses to power generation planning by the electricity industry to optimize the production and consumption of energy. Hitachi believes that it can bring about innovation in social infrastructure by combining the extensive experience it has gained through its past activities in railway, electricity, and other infrastructure with its experience from the field of IT in recent years.

The KaaS knowledge platform is an important element of this innovation and Hitachi believes that it can contribute to global issues such as energy and the environment in the future.

Hitachi intends to continue working on research and development of KaaS and other technologies with the aim of realizing a society based on knowledge creation that takes account of people, the planet, and other considerations.

REFERENCES

(1) "Undercurrent — Privacy Concern Casts Shadow over Activity Support Services," Nikkei Communications, pp. 46–50 (Feb. 1, 2009) in Japanese.
(2) NPO Japan Network Security Association, "2008 Information Security Incident Survey Report, Revised Edition 1.3," (Nov. 2009), http://www.jnsa.org/result/2008/surv/incident/index.html
(3) Y. Sato, "Trends in Research into Privacy Protection Technology for Safe and Secure Society," The Institute of Electronics, Information and Communication Engineers 17th Technical Committee on Biometric System Security (Mar. 2009) in Japanese.

ABOUT THE AUTHORS

Ryoichi Ueda
Joined Hitachi, Ltd. in 1994, and now works at the uVALUE Innovation Research Department, Systems Development Laboratory. He is currently engaged in the research and development of KaaS and distributed processing technology.

Yoshinori Sato
Joined Hitachi, Ltd. in 1994, and now works at the 7th Research Department, Systems Development Laboratory. He is currently engaged in the research and development of information security. Mr. Sato is a member of the Information Processing Society of Japan (IPSJ).

Masakatsu Mori
Joined Hitachi, Ltd. in 1994, and now works at the 2nd Research Department, Systems Development Laboratory. He is currently engaged in the research and development of cloud computing. Mr. Mori is a member of the IPSJ and The Society of Instrument and Control Engineers (SICE).

Kozo Nakamura
Joined Hitachi, Ltd. in 1977, and now works at the Hitachi Research Laboratory. He is currently engaged in the research and development of preventive maintenance systems. Mr. Nakamura is a member of The Institute of Electronics, Information and Communication Engineers (IEICE).

Nobutoshi Sagawa
Joined Hitachi, Ltd. in 1985, and now works at Research & Development Center, Hitachi Asia Ltd. He is currently engaged in the research and development of large information systems. Mr. Sagawa is a member of the IPSJ.