Two Products for the Hitachi Data Science Platform

Big Data Discovery and NX Context-based Data Management System

Today's clients are looking to IoT and big data-related business applications to utilize data as a means for creating more advanced maintenance methods, for optimizing operations, and for developing new services. These expectations are prevalent among industries with manufacturing plants and among public infrastructure providers in sectors such as rail transport, power, and gas. This article presents two products provided on the Hitachi Data Science Platform that help prepare data for utilization—Big Data Discovery and the NX Context-based Data Management System. The Hitachi Data Science Platform provides clients with comprehensive assistance for data utilization. Customers can utilize it to create a data-gathering environment or a data lake for unified data storage. It can also provide assistance with preparing data for utilization, along with AI- and BI-driven data analysis services and application development.

Takashi Tsuno
Yuuji Takamura
Toshiko Takamura

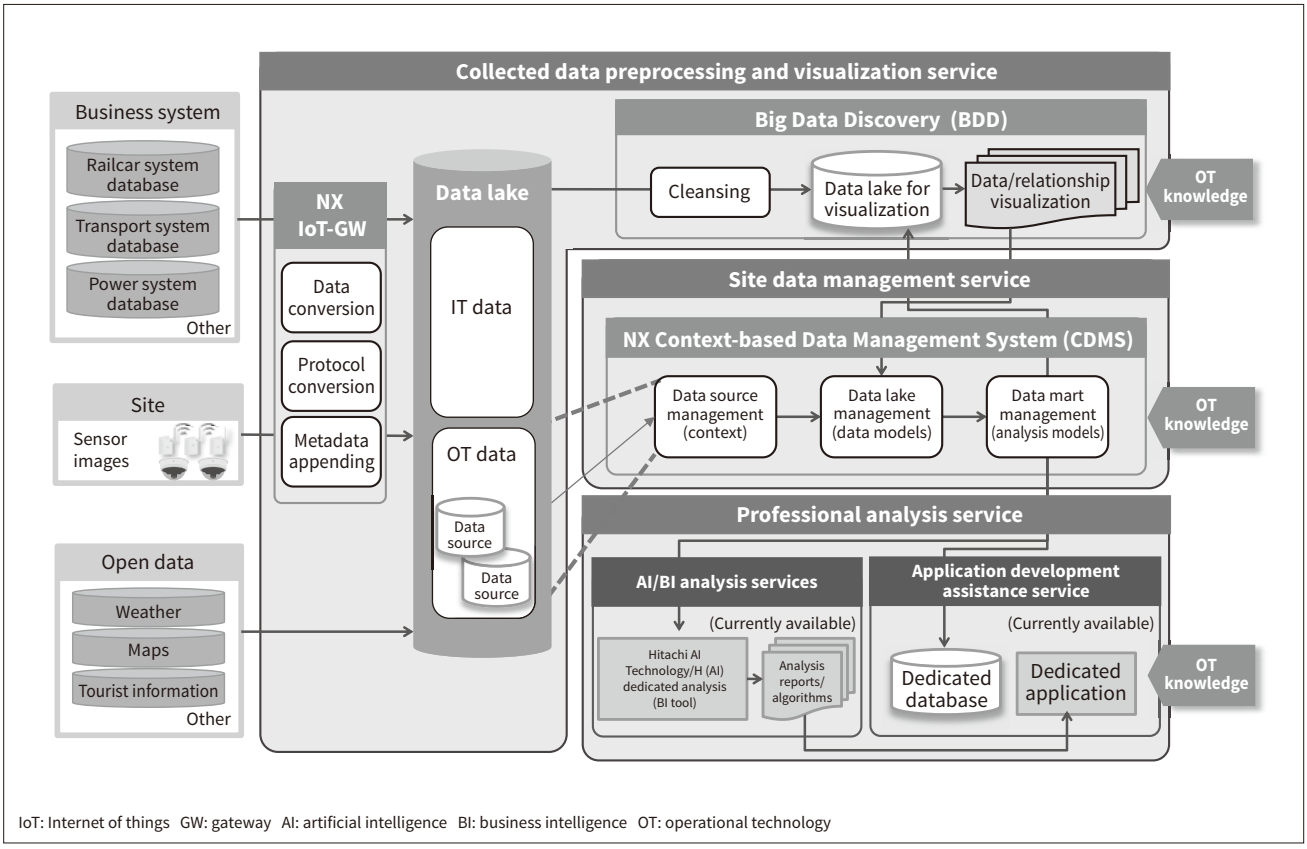
1. Introduction

Operational technology (OT) data that can be acquired from sensors mounted in various devices used at sites comes in many different formats. IT data is often managed using different data items or names for each business system even when the information it provides is the same. Utilizing these different types of data together requires preliminary preparations such as integrating data, standardizing formats, and standardizing units of measurement. Not knowing where the data they need is located, user departments that want to utilize data call on system departments to provide it. But siloed systems make data-gathering a labor-intensive process for the system departments also.

Hitachi has developed two products designed to reduce the workload needed for data preparation: Big Data Discovery (BDD) and the NX Context-based Data Management System (CDMS). BDD enables efficient extraction and creation of the required data from massive amounts of data in a wide range of formats, while CDMS enables reconfiguration and management of OT data from the user's perspective. These products assist with preliminary data preparation to enable integrated analysis and utilization of OT and IT data. BDD and CDMS are provided as collected data preprocessing/visualization and management service functions on the Hitachi Data Science Platform (DSP) (see **Figure 1**).

Figure 1 — Schematic Diagram of Hitachi Data Science Platform (DSP)

DSP provides three products/services designed to give clients comprehensive assistance with data use: (1) The collected data preprocessing and visualization service provides an NX IoT gateway function and BDD function. The NX IoT gateway function gathers sensor data from plant facilities and devices. The BDD function enables rapid merging/extraction of data for utilization by automatically visualizing data relationships among gathered OT data and IT data acquired from systems. (2) The site data management service provides an NX CDMS function that makes it easy to utilize site data by defining the meanings of analysis data items and managing models. (3) In the future, the professional analysis service will seamlessly coordinate AI and BI with DSP to provide services assisting advanced analysis and dedicated application development.



IoT: Internet of things GW: gateway AI: artificial intelligence BI: business intelligence OT: operational technology

2. BDD Overview

One solution for giving data users unrestricted utilization of data is to enable them to acquire the required data themselves. But this approach results in complex data acquisition methods and requires a solution to the issue of the extensive time needed for data acquisition processes. This section discusses the product that Hitachi has created to solve these issues, BDD.

2. 1

Issue 1: Complex Data Acquisition Methods

User departments that want to utilize data must perform complex procedures such as obtaining data usage permission for each data management department or asking for assistance from system departments. Only after completing these complex procedures can they eventually obtain the data. To solve this issue, Hitachi has devised an approach in which data is gathered in a single location and users are given the ability to acquire data themselves using a standard interface for all data types (see **Figure 2**).

2. 2

Issue 2: Extensive Time for Data Acquisition Processes

Currently, the data for utilization is often generated using processes that are individually optimized for each system. Individual processing capacities are becoming inadequate due to data bloat and the large number of output data file types. Systems that handle big data should ideally have a configuration that makes it easy to expand processing capacity to keep up with data volume growth. It should be possible to resolve this issue by creating an environment that enables rapid data file input/output using distributed

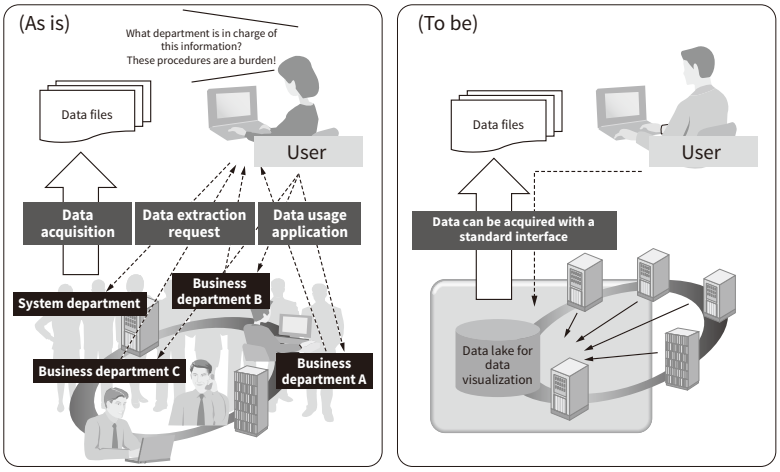


Figure 2—Improved Data Acquisition Method

User departments must currently complete complex procedures before they can eventually obtain data. To improve this situation, users need to be able to acquire data by themselves using the same standard interface for every type of data.

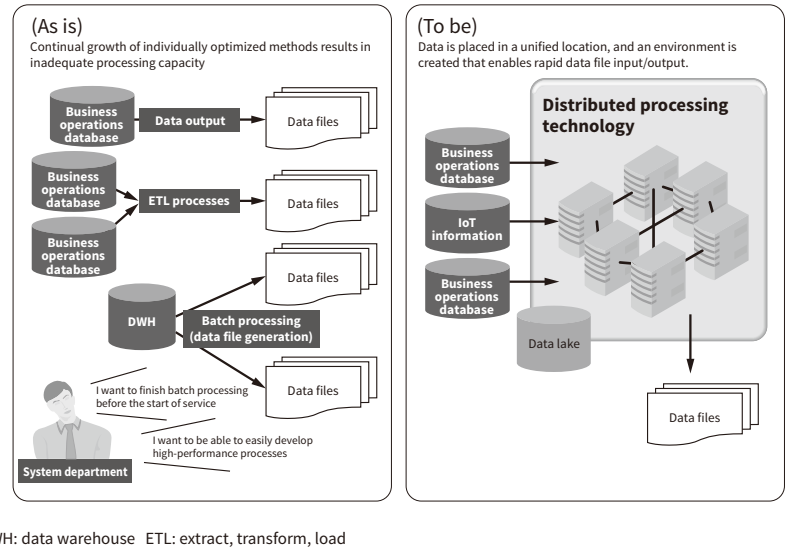
processing technology (see **Figure 3**). BDD has been designed to provide this solution. It uses the Hitachi's distributed processing platform product Hitachi Application Framework/Event Driven Computing (HAF/EDC).

2. 3 BDD Features

BDD provides a function for automatically displaying the relationship between data from different systems, and can search massive volumes of gathered data to find columns having similar names to columns of interest to the user. Its ability to easily extract data genuinely worth using enables a reduction in the workload needed to prepare data for utilization, letting the user to quickly get started with actual data utilization or another task at hand. BDD is an application for viewing, searching, linking, and creating data. Just by having the data manager load the data files, the user can check the data (data viewing), display related network diagrams (data linking), and perform other activities automatically.

(1) View

BDD makes it easy to load and view the data content of CSV-format files. Users can start by viewing the content of files on the standard interface to



DWH: data warehouse ETL: extract, transform, load

Figure 3—Improved Data Acquisition Process Time

Distributed processing technology can be used to enable rapid data file input/output, resolving the issue of extensive data acquisition process time.

determine the type of data present, letting them create a starting point for data utilization (see **Figure 4**). This function lets users view the content of large data having volumes that were previously too large to open and view as files.

(2) Search

The fuzzy search function can detect columns that have names similar to a column of interest specified as the search key. Search results are displayed in order of highest to lowest matching rate to the specified column name (see **Figure 5**). Industry-specific synonyms and related terms can also be registered in advance, such as the terms "turnout," "switch," and "point" for the rail industry.

(3) Link

Data items to be viewed side-by-side can be linked from among the searched data items. Users can extract data genuinely worth utilizing while viewing the links among the data (see **Figure 6**). The central node in **Figure 6** is the key item used to link the other data items. The first-layer nodes around it represent tables. The second-layer nodes at the outer edge represent non-key data items included in the tables. For example, if System 1 contains Table A, System 2 contains Table B, Table A contains the data item "Railcar ID," and Table B contains the data item "Inspection railcar ID," then the central node in **Figure 6** represents the side-by-side view of "Railcar ID" and "Inspection railcar ID." Functions are also provided that let the user select a second-layer data item and view the profile of the data contained in that data item as a graph such as a histogram or scatter plot.

(4) Create

Data selected for extraction can be generated as data for utilization. Data selected from multiple systems can also be merged. Created data can be used by business intelligence (BI) tools, artificial intelligence (AI), or dedicated applications (see **Figure 7**). When generating data, BDD does not fill in missing values.

Figure 4 — Viewing Data

Data is read on a viewer to check for effective data items.

Figure 5 — Searching Data

The name of the data item of interest is entered as the search key, and effective data items are selected by performing a fuzzy search spanning multiple systems.

Figure 6 — Linking Data

The user can determine whether a data item is worth using by selecting a merge key and using it to view graphical representations such as relationship diagrams, appearance frequencies, histograms, and scatter plots.

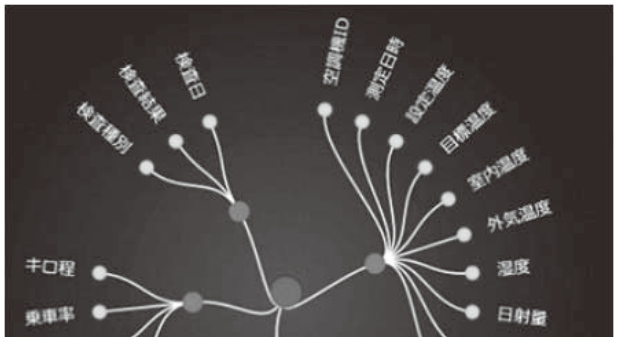
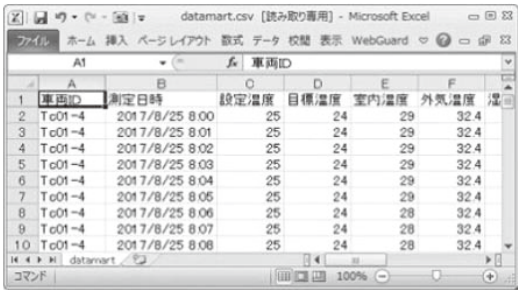


Figure 7 — Creating Data

Data determined to be worth using is output in a format enabling easy handling, and subsequently used in analysis activities and other processes.



For example, if there are missing values in a column of temperature management data, they could be filled in with average values, mode values, immediately preceding values, or by various other methods. The method used will depend on the data utilization objective. By not making up for missing values, BDD enables the actual state of the data in the company's possession to be seen. This approach lets the data user process the data as needed for the application before utilization.

3. CDMS Overview

This section describes the data management platform designed for OT data utilization in offline analysis and online applications. Many different types of equipment and sensors are found at sites. The data-gathering timing and format varies with each sensor, and sensors are managed by many different site systems. As a result, OT data has been previously handled by only experts familiar with on-site systems. Investigating the types of data present at a site is a time-consuming task for the analysts who utilize OT data, and site system operators bear the burden of gathering and providing OT data whenever analysts attempt a new analysis. This section describes CDMS, a product designed to solve these issues and to enable rapid implementation of analysis and applications.

3. 1 Management Platform Enabling Easy Utilization of Site Data

CDMS provides three data management levels consisting of data accumulation, organization, and utilization (see **Figure 8**).

(1) Data accumulation (data source management)

Site system operators can start to accumulate data by preparing data files with simple definitions and

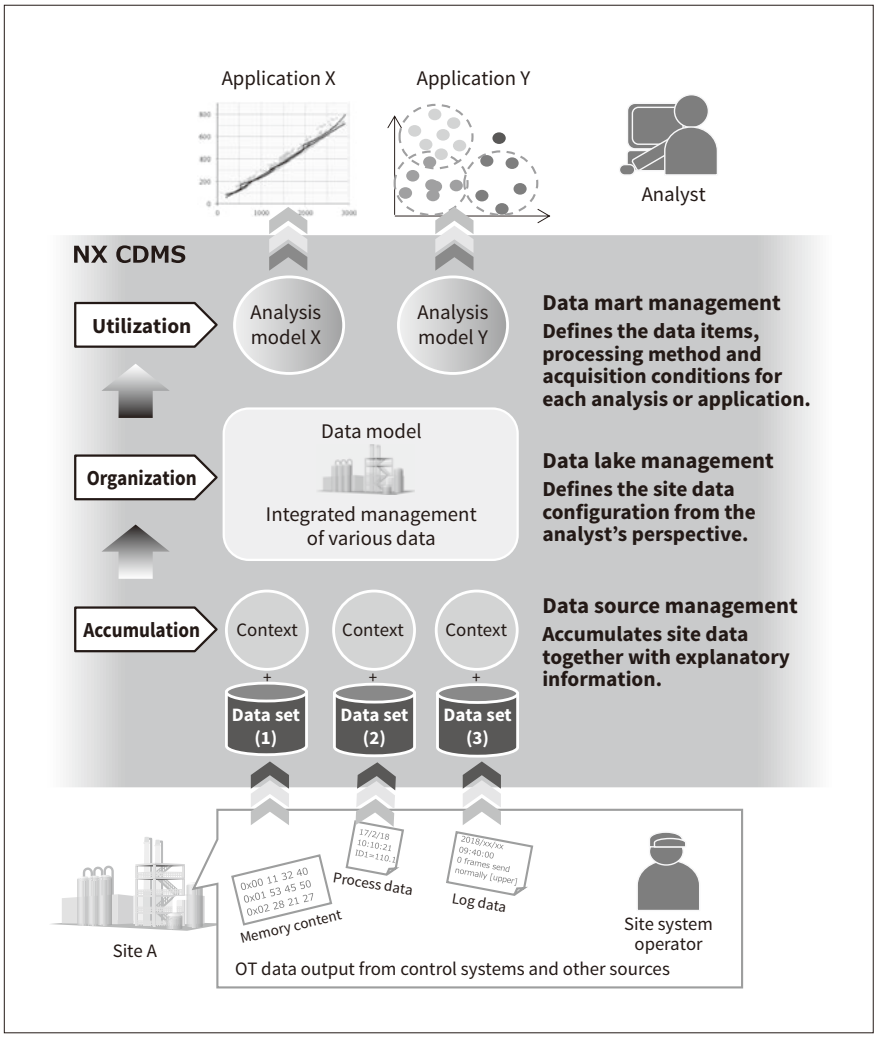


Figure 8 — CDMS Configuration

CDMS provides three data management levels consisting of data accumulation, organization, and utilization. It makes OT data handling easy even for analysts without in-depth knowledge of site systems. Additionally, it also eliminates the need for site system operators to gather and provide the data generated by each new analysis.

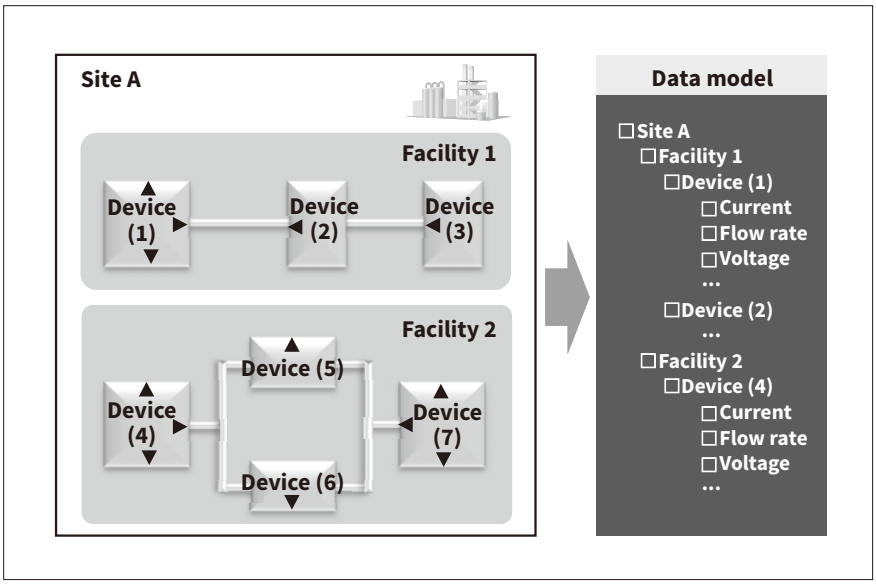


Figure 9—Data Organization
The configuration of the sensors attached to each facility and device in a building is defined as a data model. Tracing through data models lets analysts understand the type of data present at each location on site, and access the data of interest.

predetermined formats. Data sent from the site in a fixed cycle or on the timing of events is analyzed by CDMS in accordance with the definitions, and registered in data sources. Since the desired data formats will vary according to the analysis method or application, the data is accumulated in its original format without processing at this point. The explanatory information, or so-called context, which describes the unit or name of data is stored simultaneously with the sensor or device data, and that makes it easy to understand the meaning of the data itself.

(2) Data organization (data lake management)

At the data lake level, the accumulated data in data sources is redefined as data models from the perspective of analysts. The data models stand for the hierarchical configuration in which the location of each device or sensor at the site is expressed. (see **Figure 9**). Analysts can trace through these data models to find out the type of data present at each site location and access the data of interest.

(3) Data utilization (data mart management)

When utilizing data, the required data is selected from the data model. In CDMS, it is possible to define the data items, processing methods (such as time setting, unit conversion, and missing value compensation), and acquisition conditions used for each analysis method or application as an analysis model. Analysis models make it possible to retrieve the data in the same format at any time.

3. 2
Progressive Expansion of Analysis Data and Applications

As mentioned above, managing data with CDMS enables easy utilization of OT data. However, it is difficult to make a big investment in large-scale data management from the beginning. So, a data management platform is required to support gradual expansion as it operates. The required data volume will increase as the scope of analysis expands, and introducing new analysis methods and applications will create the need for data in new formats. CDMS makes it easy to add data and new analysis methods or applications by defining a data source, data lake, or data mart with the graphical user interface (GUI). The HAF/EDC, distributed processing platform, also lets users expand their central processing unit (CPU) capacity without shutting down systems, to accommodate increases in server processing volume as handled data or analysis applications increase.

4. Conclusions

Implementing the utilization of data ultimately creates new business value in the form of improved productivity and increased earnings. Clients who actively embrace the utilization of data have conducted individual analysis activities for proof of concept (PoC) purposes and other reasons, but most of them have

been one-time-only experiments. This lack of follow-up has occurred partly because the workload for preparing the data is greater than for the actual data analysis, and companies are unused to doing these activities on an ongoing basis.

BDD lets data users search and present large amounts of data to discover data worth using and output it in easily handled formats. It is hoped that the use of BDD will help reduce the data preparation workload and promote more advanced data utilization. CDMS lets data users understand the meaning of data known only at the site by assigning appropriate attribute information to it in advance. CDMS will help create benefits such as reducing accidents and stimulating the utilization of OT data. Hitachi will continue improving the functions of DSP so that it can be used to extract new value-creating data structures (with BDD), and help create ongoing value-creating data/analysis cycles in corporate activities (with CDMS).

Authors

Takashi Tsuno
Transportation Information Systems Division, Social Infrastructure Information Systems Division, Social Infrastructure Systems Business Unit, Hitachi, Ltd. *Current work and research:* Coordination of departments responsible for Big Data Discovery.

Yuuji Takamura
Social Innovation & Telecommunication Division, Social Infrastructure Information Systems Division, Social Infrastructure Systems Business Unit, Hitachi, Ltd. *Current work and research:* Development of solutions adopting distributed processing platforms.

Toshiko Takamura
Control System Platform Development Department, Control System Platform Division, Services & Platforms Business Unit, Hitachi, Ltd. *Current work and research:* Development of the NX Context-based Data Management System.