Hitachi Data Science Platform
データ分析基盤／Big Data Discovery，NX Context-based Data Management System

鉄道や電力・ガスなどの社会インフラ分野や製造プラントを有する産業分野においては，データの利活用によるメンテナンスの高度化や最適なオペレーションの実現，新規サービスの創出に向け，IoT・ビッグデータ関連の業務への応用が期待されている。
Hitachi Data Science Platformは，顧客のデータ利活用をトータルにサポートする。具体的には，データを収集する環境の構築，データを一元的に収容するデータレイクの構築，利活用のためのデータの準備支援，AIやBIを活用したデータ分析サービス，およびアプリケーションの開発などが提供可能である。
本稿では，利活用のためのデータの準備支援に該当する製品であるBig Data DiscoveryとNX Context-based Data Management Systemについて紹介する。

津野 高志 | Tsuno Takashi
高村 祐史 | Takamura Yuuji
高村 稔子 | Takamura Toshiko

1. はじめに

現場で用いられる各種機器に設置されたセンサーから取得可能なデータ［OT（Operational Technology）データ］の形式は異種混合である。また，ITデータは業務システムごとに，同じ内容が異なるデータ項目や名称で管理されていることも多い。これらのデータをまとめて利活用するためには，データの統合，形式の統一，および単位合わせなどの事前準備が必要になる。

ユーザー部門がデータを利活用したい場合，必要なデータがどこにあるか分からないため，システム部門にデータ提供を依頼することになる。しかし，システム部門にとってもシステムがサイロ化しているため，データを収集するための負荷が高くなっている。

このようなデータ準備の負荷を低減するため，膨大で多種多様な形式の情報から効率的に目的のデータを抽出・作成できる「Big Data Discovery」（以下，「BDD」と記す。）と，利活用する側の視点でOTデータの構成を再定義して管理できる「NX Context-based Data Management System」（以下，「CDMS」と記す。）を新たに開発した。これにより，OTデータとITデータの統合的な分析・利活用の事前準備を支援する。

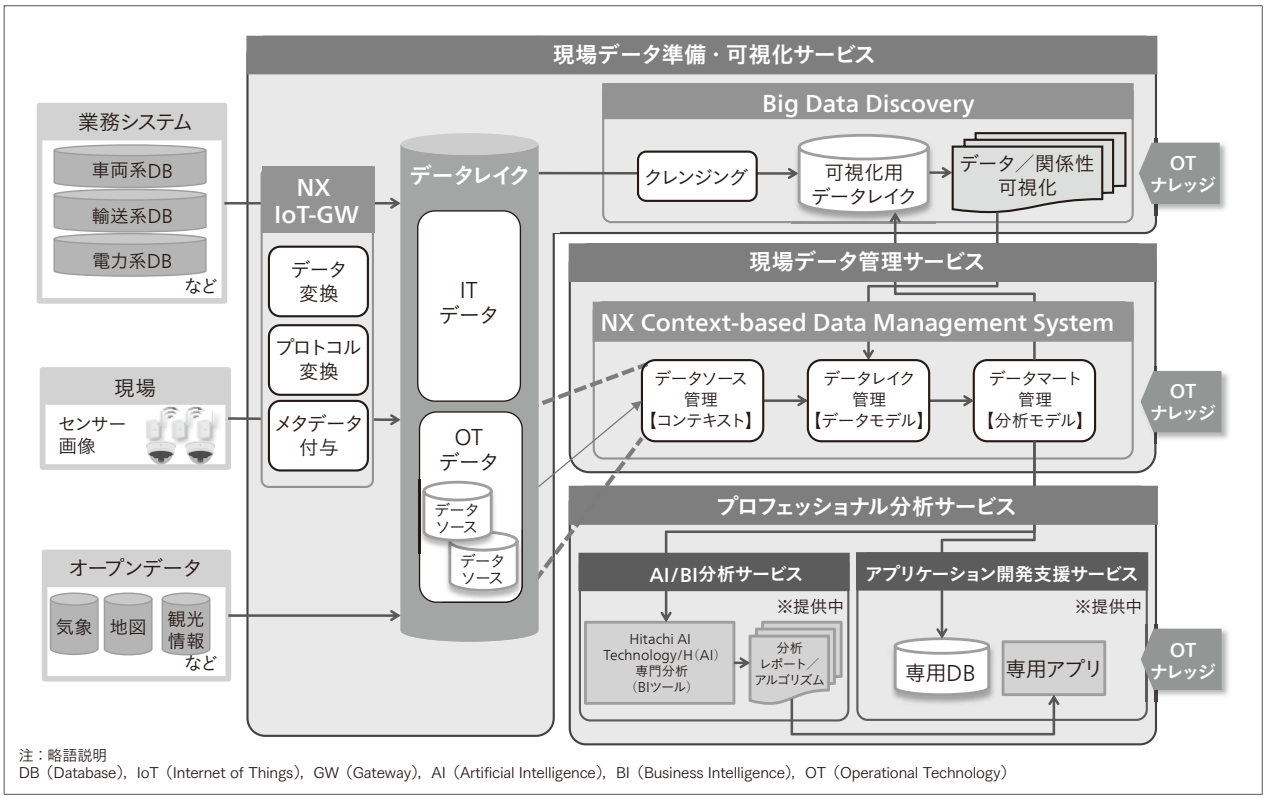
なお，BDDとCDMSは「Hitachi Data Science Platform」（以下，「DSP」と記す。）上の現場データ準備・可視化／管理サービス機能として提供する（**図1**参照）。

2. BDDの概要

データを利活用する人が，データを自由に活用できるようにするためには，自分で必要なデータを取得できるようにしていくことが解決策の一つだと考える。そのた

図1｜DSP（Hitachi Data Science Platform）の全体図

DSPは，顧客のデータ利活用をトータルにサポートするために3つのサービス／製品を提供する。現場データ準備・可視化サービスでは，プラント設備・機器などからセンサーデータを収集するNX IoT-GW機能と，収集されたOTデータやシステムから取得したITデータを対象に，データの関係性を自動で可視化して，利活用対象とすべきデータを高速に結合／抽出可能にするBDD機能を提供する。現場データ管理サービスでは，分析対象のデータを意味付け，モデル管理することで，現場データの利活用を容易にするNX-CDMS機能を提供する。また，プロフェッショナル分析サービスでは，DSPとAIやBIをシームレスに連携することで高度な分析を支援するサービスや，専用のアプリ開発を今後提供していく。



注：略語説明
DB（Database），IoT（Internet of Things），GW（Gateway），AI（Artificial Intelligence），BI（Business Intelligence），OT（Operational Technology）

めには，データを取得する手段が煩雑，データ取得処理の時間が長大といった課題を解決する必要がある。本章では，これらの課題を解決するための製品であるBDDについて説明する。

2.1
課題1：データを取得する手段が煩雑

ユーザー部門がデータ利活用をしたい場合，データ管理部門ごとにデータ使用許可を取ったり，システム部門に依頼したりするなど，手続きが煩雑な状態である。この煩雑な手続きなどを通して，ようやくデータを入手できるようになるという現状の課題を解決するためには，データを1か所に集めて，どのようなデータでも統一的なインタフェースで，ユーザー自身でデータを取得することができるようになればよいと考える（**図2**参照）。

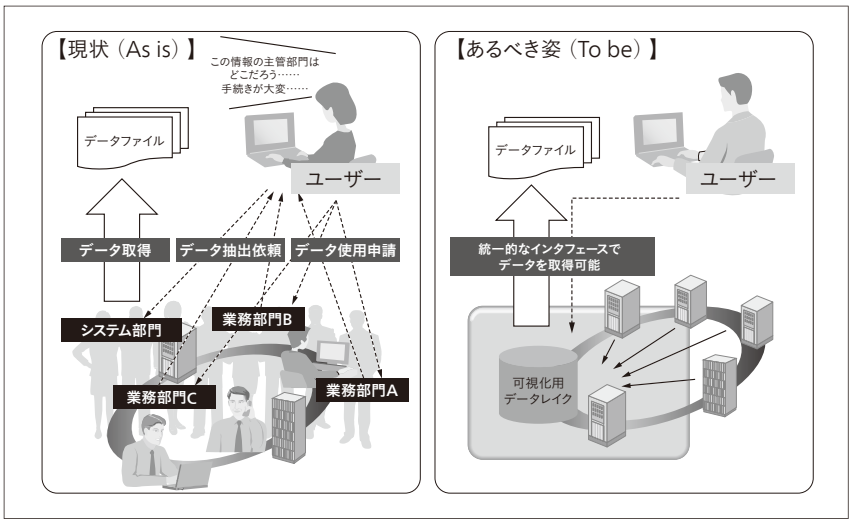
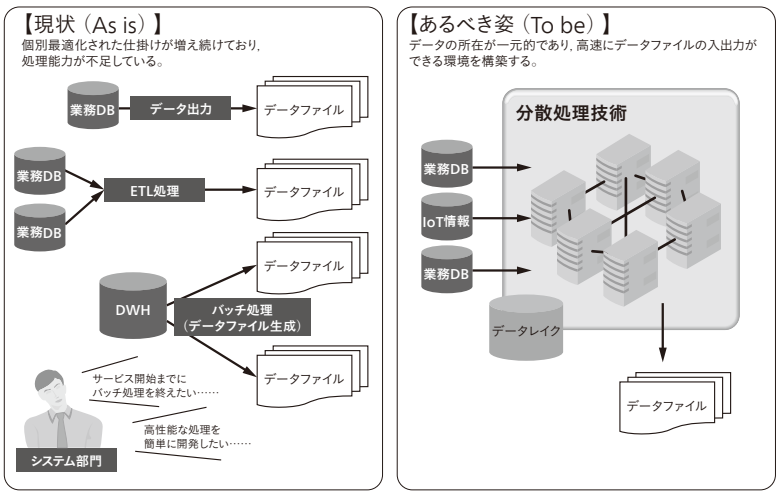


図2｜データ取得手段の改善

煩雑な手続きなどを通して，ようやくデータを入手できるようになるという現状を改善するには，どのようなデータでも統一的なインタフェースで，ユーザー自身でデータを取得することが可能になればよい。



注：略語説明
DWH（Data Warehouse），ETL（Extract, Transform, Load）

図3｜データ取得処理時間の改善
分散処理技術を使って高速にデータファイルの入出力を可能にすることで，「データ取得処理の時間が長大」という課題を解決する。

2.2
課題2：データ取得処理の時間が長大

現状は，利活用のためのデータの生成は，システムごとに個別最適化された処理にて行っている場合が多い。データ容量の肥大化，出力するデータファイルの種類が多くなっているといった事情から，個々の処理能力が不足してきている。

ビッグデータを扱うシステムは，データ容量の増大に追随できるように処理能力を拡張しやすい構成が好ましい。分散処理技術を使って，高速にデータファイルの入出力ができる環境を構築することで，この課題を解決できると考える（**図3**参照）。BDDは，そのための製品として,日立の分散処理基盤製品であるHitachi Application Framework / Event Driven Computing（HAF/EDC）を使用している。

2.3
BDDの特徴

BDDは，異なるシステム間のデータの関係性を自動で表示する機能を備えており，収集した膨大なデータの中から，自分が着目したカラムと類似した名称のカラムを見つけ出すことが可能である。真に利活用対象とすべきデータを容易に抽出できるため，利活用のためのデータの準備に要する負荷を軽減し，本来の利活用作業などに迅速に取りかかることが可能になる。

BDDは，データを「見る」，「探す」，「つなぐ」，「作る」アプリケーションである。データ管理者がデータファイルを投入するだけで，利用者によるデータ確認（データを「見る」）や関係ネットワーク図表示（データを「つなぐ」）などの活動が自動的にできるようになる。

(1)「見る」

CSV（Comma-separated Values）形式のファイルをBDDに投入でき，データの中身を簡単に参照することができる。まずはどのようなデータが存在するのかを統一的なインタフェースで参照できるようにすることで，データ利活用の足がかりを作ることができる（**図4**参照）。

この機能により，ファイルベースでは開くことができないため参照できなかった容量の大きなデータの中身を見ることが可能になる。

(2)「探す」

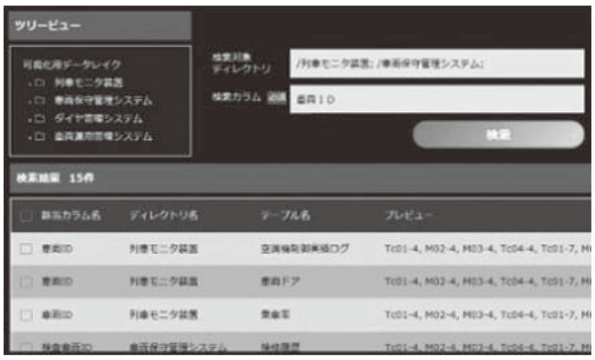
「あいまい検索」機能で，自分が着目したカラムをキーに検索することで，該当カラムと類似した名称のカラムを検出できる。検索結果はカラム名の一致度が高い順に表示する（**図5**参照）。

また，業界特有の類似語をあらかじめ登録することも可能である。例えば，鉄道業界における，転てつ機，スイッチ，ポイントといった用語である。

図4｜データを「見る」
ビューアを使用してデータを見ながら有効なデータ項目を確認する。

図5｜**データを「探す」**

着目したデータ項目の名称をキーに，複数システムを横断したあいまい検索により有効なデータ項目を選定する。



(3)「つなぐ」

検索したデータ項目の中から，同一視したいデータ項目をつなぐことができる。利用者はデータのつながりを見ながら，真に利活用対象とすべきデータを抽出することが可能である（**図6**参照）。

同図の中央の●は，データ項目をつなぐためのキー項目である。その周囲にある1階層目の●はテーブルである。さらに外側の2階層目の●はキー項目以外にテーブルに含まれていたデータ項目となる。

例えば，システム1にテーブルAがあり，システム2にテーブルBがあるとする。テーブルAには車両IDというデータ項目があり，テーブルBには検査車両IDというデータ項目があるとする。この車両IDと検査車両IDを同一視したものが同図の中央の●である。

2階層目のデータ項目を選択し，そのデータ項目に含まれるデータのプロフィールを確認するためのヒストグラム，散布図などの機能も備えている。

(4)「作る」

抽出対象として選択したデータを，利活用対象のデータとして生成することができる。また，複数のシステムから選択したデータを結合することもできる。作成したデータは各種BI（Business Intelligence）ツールやAI（Artificial Intelligence），専門アプリケーションで利用が可能である（**図7**参照）。

なお，データを生成するにあたって，BDDはデータの欠損値を埋めることはしない。例えば，温度を管理するカラムのデータに欠損値があった場合，そこを平均値で埋める，最頻値で埋める，直前の値で埋めるなどさまざまな対処方法が考えられるが，どのように対処するかは，そのデータの利活用の目的に依存する。BDDは企業の持つデータは実際にはどういう状態かを可視化するために，欠損値の補完はしない。そのため，データを利活用する人が目的に応じた対処をした後で，データを利用することになる。

図6｜**データを「つなぐ」**

選定した結合キーで関係図や出現頻度，ヒストグラム，散布図などを確認し，利用価値があるデータ項目かどうかを判断する。

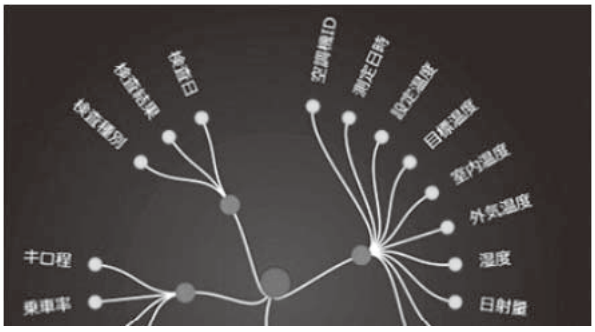
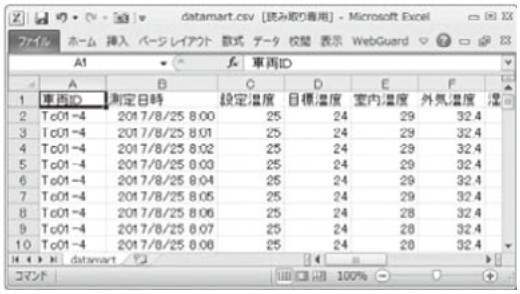


図7｜**データを「作る」**

利用価値があると判断したデータを扱いやすい形で出力し，その後，分析活動などで利用する。



3. CDMSの概要

次に，OTデータのオフライン分析やオンラインアプリケーションへの適用に向けたデータ管理基盤について述べる。現場にある機器やセンサーは多種多様である。データを収集するタイミングや形式はセンサーごとに異なり，またセンサーを管理している現場システムもさまざまである。このためOTデータは，現場システムに精通する専門家にしか取り扱えなかった。OTデータを使用する分析者は，現場にどのようなデータがあるのか調査するのに時間がかかる。また現場システム運用者は，分析者が新しい分析を試みるたびに，OTデータを集めて提供するのが負担となっている。本章では，これらの課題を解決し，分析やアプリケーションを短期間で実現するための製品であるCDMSについて説明する。

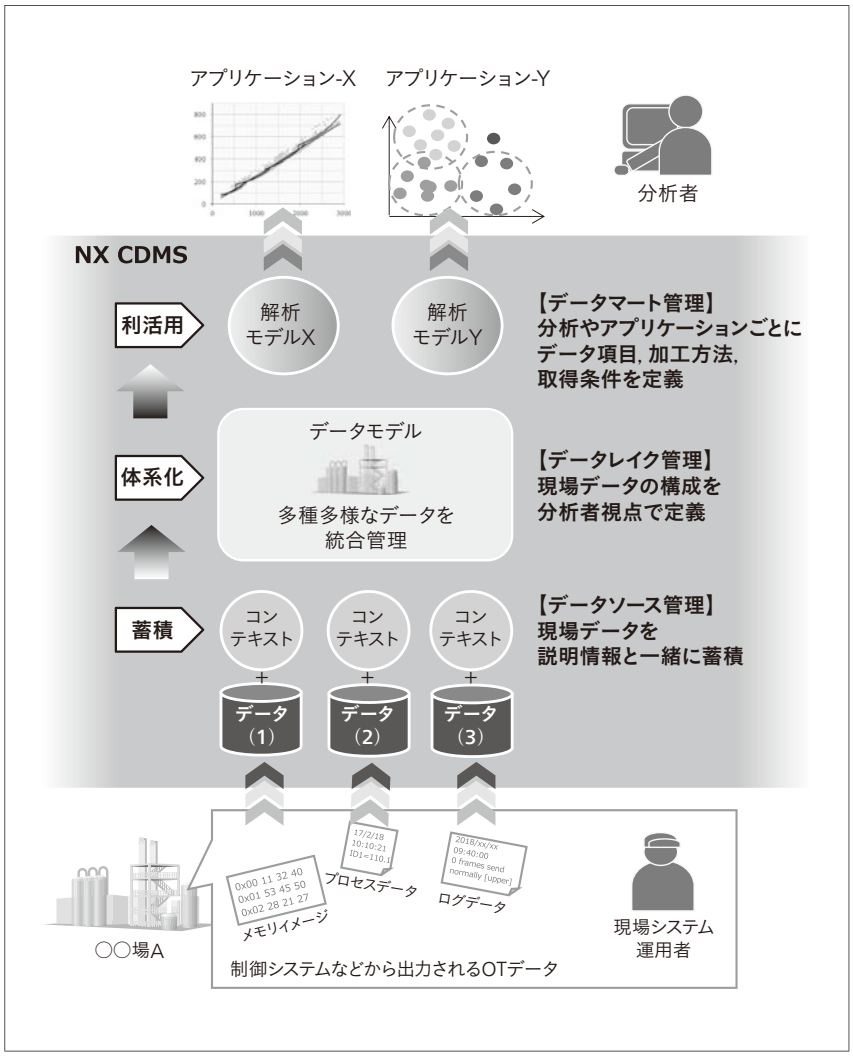


図8｜CDMSの構成

CDMSでは，データを「蓄積」,「体系化」,「利活用」の3段階で管理する。これにより，現場システムに精通していない分析者でもOTデータを容易に扱えるようになる。また，現場システム運用者は，新しい分析のたびに発生するデータの収集・提供が不要になる。

3.1
現場データを手軽に利用できる管理基盤

CDMSは，データを「蓄積」,「体系化」,「利活用」の3段階で管理している（**図8**参照）。

(1) データの蓄積「データソース管理」

現場システム運用者は，簡単な定義と所定の形式のデータファイルを準備することでデータの蓄積を開始できる。CDMSは，一定の周期やイベントごとに現場から送信されてくるデータを定義に従って解析し，データソースに登録する。分析方法やアプリケーションによって欲しいデータの形は異なるため，ここではデータの加工はせず，そのままの形で保管する。機器やセンサーなどのデータに，単位や名称などデータを説明する「コンテキスト」情報をひも付けて管理することで，データの意味を分かりやすくする。

(2) データの体系化「データレイク管理」

データレイクでは，蓄積されたデータを分析者の視点で「データモデル」として定義する。ここでいうデータモデルとは，各機器やセンサーが，どの建屋の中のどの設備・装置に付いているものなのか，その構成を階層で整理したものである（**図9**参照）。分析者は，このデータモデルをたどることで，現場のどこにどのようなデータがあるのか理解でき，目的のデータにアクセスすることができる。

(3) データの利活用「データマート管理」

データを利用するときは，データモデルから必要なデータを選択する。CDMSでは，分析方法やアプリケーションごとに使用するデータ項目，時刻合わせ／単位変換／欠測値補完などの加工方法，取得条件を「解析モデル」として定義しておくことができる。解析モデルを用いることで，いつでも同じ形でデータを取り出すことができるようになる。

3.2
分析対象データやアプリの段階的な拡張

CDMSでデータを管理することで，OTデータの利用が簡単にできる。しかし，初めからすべてのデータを管理するための大きな投資は難しい。データ管理基盤には，

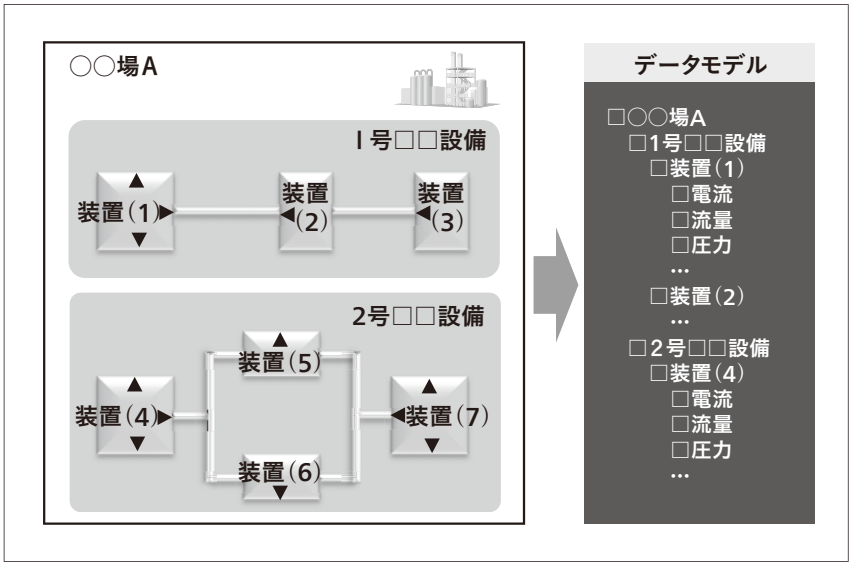


図9｜データの体系化

建屋の各設備・装置にどのようなセンサーが付いているのか，その構成を「データモデル」として定義する。分析者は，このデータモデルをたどることで，現場のどこにどのようなデータがあるのかを理解でき，目的のデータにアクセスできる。

運用をしながら段階的に拡張できることが要求される。

分析対象範囲を広げていくと，必要なデータが増加していく。また，新しい分析方法やアプリケーションを導入すると，今までとは異なる形のデータが必要になる。CDMSではGUI（Graphical User Interface）でデータソース，データレイク，データマートの定義をすることで，データの追加，新しい分析方法やアプリケーションの追加に容易に対応できる。

取り扱うデータの増加や分析アプリケーションの増加に伴い，サーバの処理量も増加する。分散処理基盤であるHAF/EDCを活用することで，CPU（Central Processing Unit）の能力拡張もシステムを停止することなく行うことができる。

4. おわりに

生産性向上や収益増加といった新たな事業価値はデータ利活用を具現化した先にある。データ利活用に積極的な顧客は，PoC（Proof of Concept）の意味も込めて，個別のデータ分析活動を推進してきたが，その多くが単発的な活動で終わってしまった。その原因の一つに，本来のデータ分析の負荷よりもデータを準備する負荷の方が大きく，企業の持続的な活動としてなじまない点があったことは否めない。

データを利活用する人が，多数のデータの中から検索や可視化を行い利用価値のあるデータを発見し，扱いやすいデータを出力できるBDDを使って，データの準備の負荷を改善し，より高度なデータ利活用を実現することに貢献したい。

また，CDMSを使って，現場でしか分からないデータについて，あらかじめ適切な属性情報を付与しておくことで，データを利活用する側にもデータの意味が分かるようにし，OTデータの利活用が活発化し，事故数の低減などを実現することに貢献したい。

DSPを活用することで，BDDで新たな価値化データ構造を抽出して，CDMSで企業活動における継続的な価値化データ・分析サイクルの確立に貢献できるように，引き続きDSPの機能強化を図っていく。

執筆者紹介

津野 高志
日立製作所 社会ビジネスユニット 社会システム事業部
交通デジタルソリューションセンタ 所属
現在，BDDの製品責任部署の取りまとめに従事

高村 祐史
日立製作所 社会ビジネスユニット 社会システム事業部
社会・通信ソリューション本部 所属
現在，分散処理基盤を適用したソリューション開発に従事

高村 稔子
日立製作所 サービス＆プラットフォームビジネスユニット
制御プラットフォーム統括本部 制御プラットフォーム開発部 所属
現在，CDMSの開発に従事