組織を越えたデータ利活用を安全・便利にする次世代データハブ

大村 圭[†1]　ジェイ ホンジェ[†1]
片山 翔子[†1]　河井 彩公子[†1]
柏木 啓一郎[†1]　馬越 健治[†2]
除補 由紀子[†1]　木村 達郎[†1]

NTTソフトウェアイノベーションセンタ[†1]
NTT社会情報研究所[†2]

経済的発展と社会課題解決の両立が期待されるデータ駆動型社会においては，企業や組織を越えたさまざまなデータの利活用が必須です．しかし，機微なデータやアルゴリズムの取り扱いや，遍在する多種多様なデータ群からの所望のデータの発見や取得に関して多くの課題が存在し，広く実践されるに至っていません．本稿では，それら課題を解決し，組織を越えたデータ利活用を安全・便利にする次世代データハブと主要な技術を紹介します．

次世代データハブとは

スマートシティや企業横断型DXなどの例にみられるように，企業や組織の枠を越えたデータの流通と利活用による新たな価値の創出や社会的課題の解決に向けた取り組みが始まりつつあります．しかし，このような取り組みを一部の限定的なものではなく，社会全体に拡大していくには以下のような課題があります．

- それぞれの企業で個別に収集，管理されている膨大なデータの中から目的に合致するデータを見つけ出すことや，必要時に迅速に取得することが困難である．
- 一方，データを提供する立場からすると，自身の貴重なデータの利用可否を適切に制御し，また，その流通範囲や利用実績を把握することが困難である．
- 特に希少性の高いデータや重要なノウハウの詰まった分析アルゴリズムを他の企業に利用させるには，目的外利用などによる機密情報漏洩の懸念がある．

これら課題の解決に向け，私たちは次世代データハブの研究開発に取り組んでいます．次世代データハブは，提供者によるデータに対するガバナンスを維持しつつ，多拠点に遍在するデータから利用者が必要なデータを迅速に効率良く入手し利用できるようにするデータ流通の基盤であり，企業や組織の枠を越えたデータの利活用を安全で便利にするものです．

次世代データハブの主要な構成要素は以下の3つです（**図1**）．

① 複数企業のデータを仮想的に統合し，効率的なデータ検索・取得を可能にする「仮想データレイク」
② 多拠点間での効率的なデータ送受信を可能にする「データブローカー」
③ 企業間でデータやアルゴリズムを互いに秘匿したまま実行可能にする「データサンドボックス」

以下，それぞれが解決する課題とアプローチを説明します．

仮想データレイク

複数の組織や拠点に遍在するデータを活用する場合，従来モデルでは，**図2**(a)に示すように各拠点で生成されたデータを単一の拠点に集約して1つの巨大なデータレイクをつくり，各利用者はそこにアクセスすることでデータ活用を行ってきました．しかし，このモデルでは，データ利用者が実際に利用するデータはごく一部であっても全量コピーが必要となるなどの一般的な問題に加え，データ提供者の視点からは自身のデータのコピーが大量に生成されデータへのガバナンスが効かなくなるといった問題が生じ，企業間でデータを流通させ，さまざまな分析・解析をすることが困難でした．

そこで私たちは，各拠点に遍在するデータを単一拠点に集約することなく活用可能とするためのデータ基盤技術を仮想データレイクと呼び，研究開発に取り組んでいます．

仮想データレイクは，図2(b)に示すとおりデータそのものではなくメタデータ*1を収集することで，遍在するデータを仮想的に集約・一元化し，データ利用者がオンデマンドに必要なデータのみを効率良く取得して分析や解析処理に活用することを可能とします．また，データ提供者からするとマスタとなるデータを常に自拠点で管理し，要求に応じて仮想データレイクを介して利用者にデータを提供することで，自身のデータに対するガバナンスを維持することが容易となります．

この仮想データレイクの実現に向け，2つの観点からの問題の解決に取り組んでいます．1番目は，さまざまな取得経緯により形式や品質の異なる膨大なデータの中からデータ利用者の目的達成に必要なデータをどのように効率的に発見し，活用につなげるかというデータ発見・活用の観点，2番目は，各拠点で日々生成され増加していく遍在データをどのように効率良く管理するか，また，データ利用者が常に利用可能な状態とし適切なタイミングで手元に届けるかというデータ管理・配信の観点です．以降ではそれぞれの観点での問題解決の取り組みについて説明します．

*1　メタデータ：本稿では，データの所在や作成者，形式などの「データを説明するためのデータ」全般のことをメタデータと呼びます．

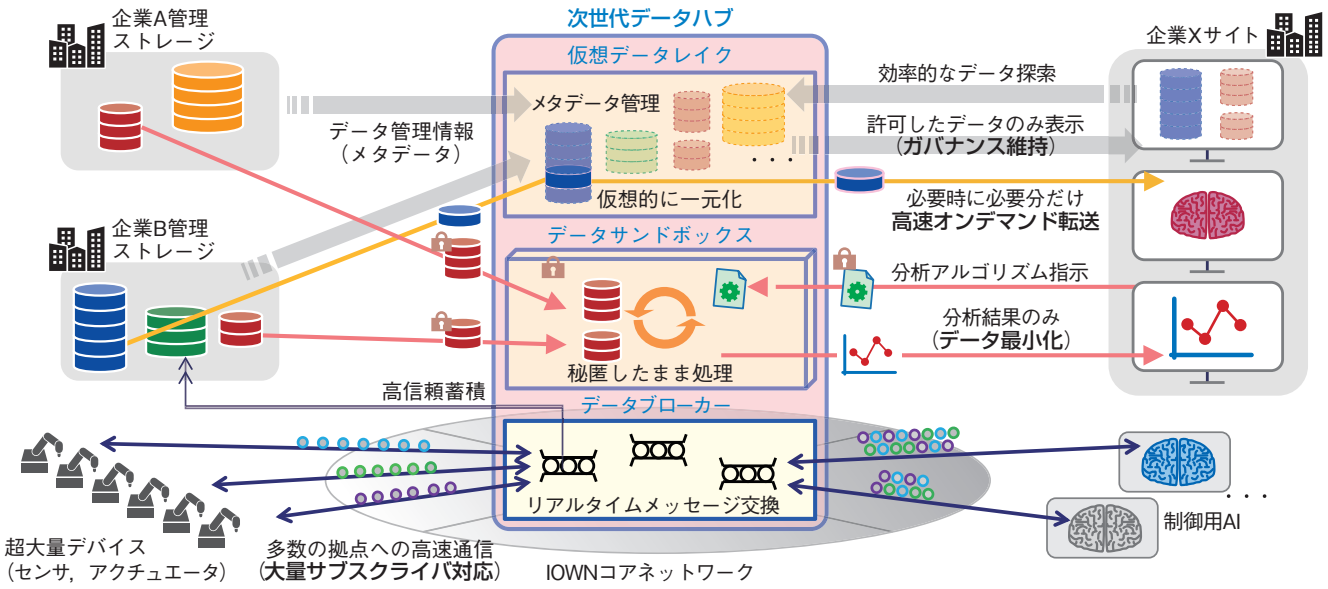

図1　次世代データハブの全体像

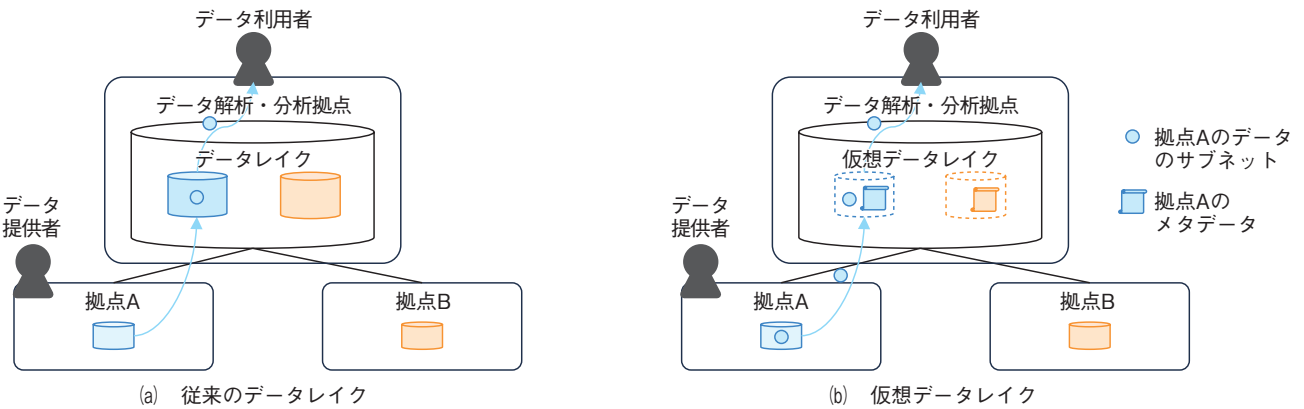

図2　データ活用モデル

■データ発見・活用

膨大なデータの中からデータ利用者が必要とするデータを発見できるようにするために，データを詳しく説明する“メタデータ”を一定の規則に従って統一的に付与し，管理します．

メタデータには，データの意味情報のようにデータ提供者によってあらかじめ付与されるものもあれば，データの形式や品質情報のように自動付与されるものもあります．これらを用いたさまざまな条件での柔軟な検索を可能とし，データ利用者の目的達成に必要なデータの絞り込みを容易にします．

また，データ間の関連性や，プロビナンスデータと呼ばれるデータの出自から流通経路や加工プロセスに至る情報もメタデータとして付与し，管理します．前者により，データ利用者が曖昧な手掛かりしかない場合でも関連したデータをたどることで本当に必要なデータを発見しやすくなります．また，後者により，データに不適切な加工がなされていないことや出自が不明な怪しいデータではないことを確認でき，データの信頼性を判断できるようになります．

■データ管理・配信

遍在するデータを集約することなく，遍在した状態のまま効率良く管理し，オンデマンドに取得できるようにするには，いくつかの課題があります．本稿では特に，各拠点で日々生成，更新されていくデータの最新状態を遠隔からどのように効率良く把握し管理するか，データの要求から返却までのレスポンスタイムを高速化し単一データレイクに集約したときのそれにどのように近づけるか，といった課題への取り組みについて述べます．

すべての拠点のデータを利用者がいつでも活用できる状態とするために，各拠点でのデータの作成・更新・削除といったイベントを低遅延に収集しメタデータとして管理します．こうした最新の状態のメタデータを用いることで，利用者からは手元にデータが存在するかのように，ファイルの一覧情報を表示したり，要求を出してその中身を取得したりすることが可能となります．

また，利用者がデータを要求してから返却されるまでのレスポンスを実用上問題のない長さにするために，データの差分配信を行えるデータフォーマットや，仮想データレイク内でのキャッシュ機構を導入します．これにより機械学習のように同じデータを繰り返し活用する処理においては，データの転送に伴う処理時間の増大を抑えることが可能となります．また，データへのアクセスのパターンや頻度に応じて動的にデータをキャッシュするプリフェッチ機能や各拠点から仮想データレイクへPUSH型の配信を行う機能を整備し，データ利用者や提供者が自身の要件に応じたアルゴリズムをプラガブルに適用できるように，設計・開発を進めています．

データブローカー

近年，スマートファクトリーやコネクテッドカーなどの分野では，数万～数10万台のセンサや車両などの端末がネットワークに接続され，監視や制御に必要なメッセージの送受信を行っています．正確な監視や制御のためには，端末からのデータ収集と端末へのフィードバック伝達をごく短時間で確実に行う必要があります．私たちは，このような大量の端末との間での低遅延かつ高信頼なメッセージ交換を可能とする，新たなデータブローカー技術の開発に取り組んでいます．

従来のデータブローカー技術では，大量端末への同報配信を効率良く行うことのみを重視してメッセージの永続化や再送制御を行わず信頼性に欠ける設計，もしくは，確実な配信のみを重視して大量端末に対応できない設計のいずれかとなっており，前述した要件の双方を同時に満足することができません．これらを両立するうえでの課題の１つとして，端末が受信できなかったメッセージの再送処理があります．再送のためには，端末１つひとつについてメッセージの配信状態を記録する必要がありますが，端末が大量に接続される場合，その状態管理コストは非常に大きく，リソース不足による転送遅延が生じ得ます．私たちは送受信プロトコルの見直しや状態管理アルゴリズムの改善によりこの課題の解決を試みています（**図３**）．

こうした課題解決の積み重ねにより，大量端末との低遅延かつ高信頼なメッセージ交換を可能とする新しいデータブローカーを実現していきます．

データサンドボックス

■背景

前述のように，異なる企業がそれぞれ保有する希少性の高いデータやアルゴリズムを掛け合わせることで新たな価値を生むことが見込まれているにもかかわらず，その漏洩への懸念から，そうした企業間の連携はほとんど行われていません．

データやアルゴリズムを互いに開示せずに掛け合わせる単純な方法とし

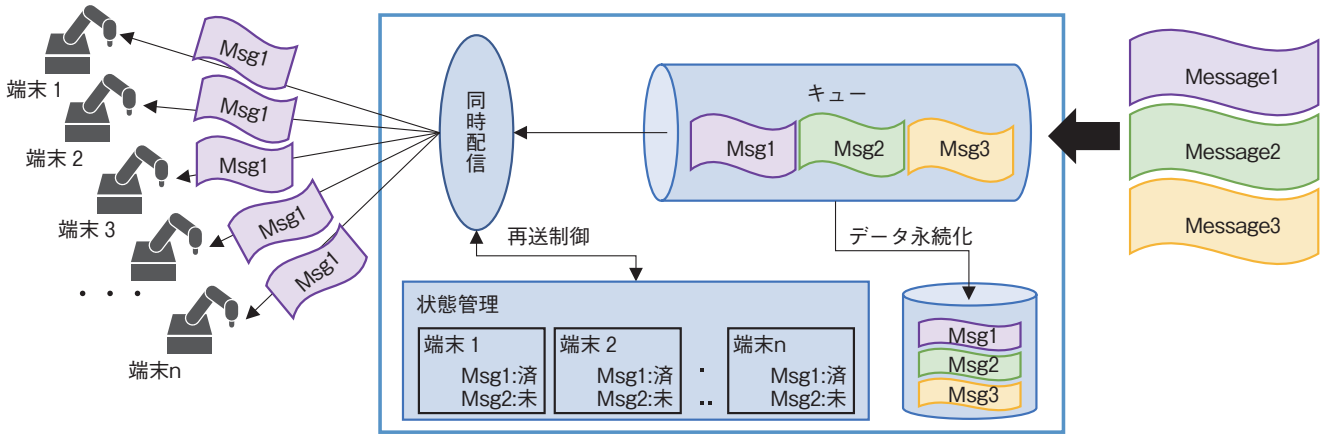

図 3　データブローカーの再送・永続化機構イメージ

て，「データやアルゴリズムを信頼できる第三者（プラットフォーム事業者）に預けて計算した結果のみを返してもらう」というものが考えられます．この方法では，データやアルゴリズムを互いに開示する必要はなくなりますが，プラットフォーム事業者には開示してしまうことになります．

私たちは，プラットフォーム事業者が計算を代行するモデルをベースとしつつ，プラットフォーム事業者にもその内容を秘匿したままデータとアルゴリズムを掛け合わせることが可能な「データサンドボックス技術」の研究開発を進めています．本技術により以下のようなデータやアルゴリズムの利活用の実現をめざしています．

- 競合各社がデータを持ち寄って統合して処理し，その結果を各社で共有する（ただし，各自のデータは相手やプラットフォーム事業者に開示されない）．
- 自社の貴重なデータを他社が有する秘伝の分析プログラムで分析し，その結果を得る（ただし，元のデータや分析プログラムは相手やプラットフォーム事業者に開示されない）．

■技術課題と解決アプローチ

これらを実現するには，主に以下に示す課題の解決が必要です．

① なりすましによるアルゴリズムの不正実行，それによる元のデータや結果の不正取得
② 誤ったアルゴリズムの実行による元のデータや結果の流出
③ プラットフォーム事業者によるデータやアルゴリズムの不正取得

データサンドボックスでは，これら課題を下記技術の組み合わせにより解決します（**図 4**）．

■技術 1 ：認証と合意に基づくアクセス制御

データサンドボックスは利用するユーザ（データやアルゴリズムの提供者，結果データの利用者）を認証し，なりすましを防止します．また，元データやアルゴリズムおよび結果データに対するアクセスをユーザどうしが事前に合意した内容（データ利用ポリシー）にのっとって制御し，アルゴリズムの不正実行やデータの不正取得を防ぎます．

■技術 2 ：揮発性のある独立した実行環境の生成

データサンドボックスはユーザどうしが合意したデータ利用ポリシーごとに独立した実行環境を生成します．また，この実行環境から外部へのアクセスを制限し，アルゴリズムによる利用ポリシーに反したデータの外部への持ち出しを防止します．さらに，アルゴリズムの処理完了と同時にこの実行環境を消去し，データやアルゴリズムの流出を防止します．

＊2　TEE：OSの管理権限を持つユーザによるメモリ参照を防止するため，CPUがメモリ領域を暗号化して利用するように構成された隔離実行環境のこと．これまではモバイル端末や組込み系機器を中心に利用されていましたが，近年ではIntelやAMDなどのサーバ用CPUにも多く搭載され始めています．

＊3　リモートアテステーション：ユーザがTEEの真正性を確認する手段としてCPUベンダが提供する機能．ユーザはTEEの構成情報を取得し，これをCPUベンダがインターネット上で提供するリモートアテステーションサービスを用いて検証することで，そのTEEがCPUベンダの提供した機能を利用して作成され，改ざんされていないことを確認できます．

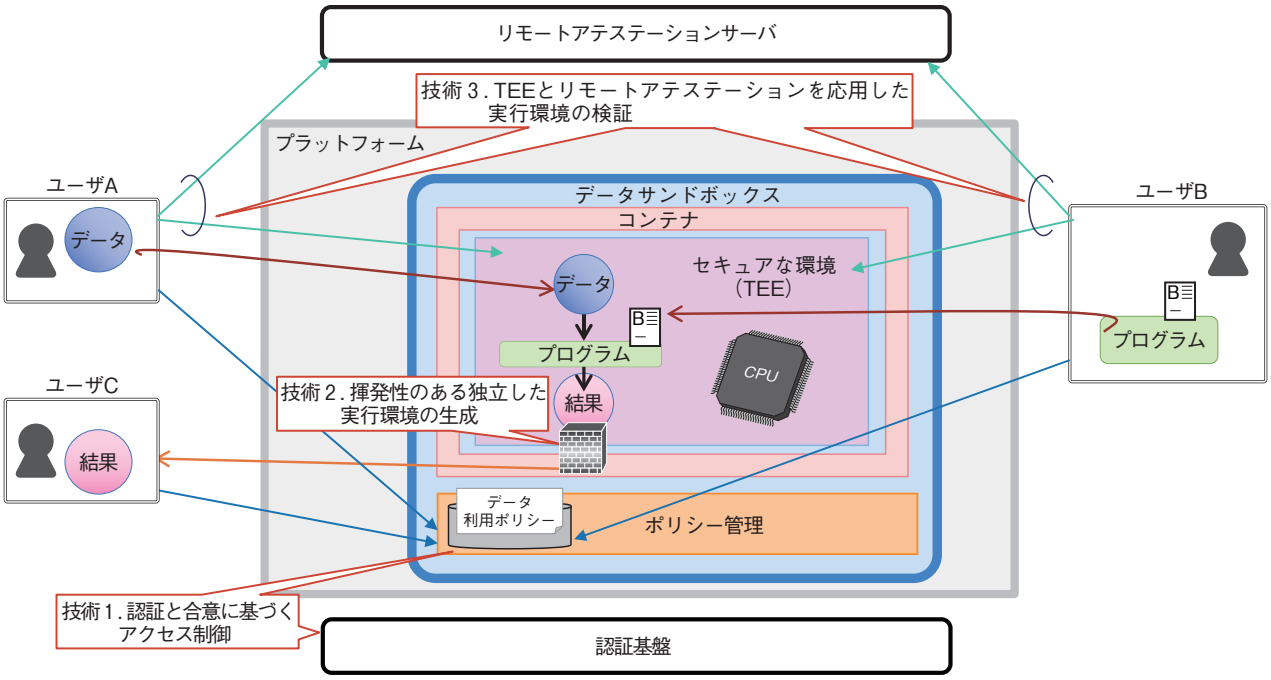

図４　データサンドボックスの技術構成

■技術３：TEEとリモートアテステーションを応用した実行環境の検証

プラットフォーム事業者など実行環境を運用，管理する者によるデータやアルゴリズムの参照を防止する既存の仕組みとしてTEE (Trusted Execution Environment)[*2]とリモートアテステーション[*3]があります．データサンドボックスは，これらを応用することで，「実行環境に取り込んだデータやアルゴリズムがデータ利用ポリシーで合意されたものであること」，および「利用している実行環境がTEEを用いて生成され，データやアルゴリズムが秘匿化されていること」をユーザ自身により検証可能としています．これにより，プラットフォーム事業者によるデータやアルゴリズムの参照を防ぎつつ，データとアルゴリズムを掛け合わせて結果を得ることを可能としています．

今後に向けて

本稿では，私たちが研究，実用化を進めている次世代データハブの主要な構成要素である「仮想データレイク」「データブローカー」「データサンドボックス」を詳しく紹介しました．次世代データハブにより，安心・安全かつ効率的なデータ流通が可能になり，これまで困難とされてきた企業や組織の枠を越えた機密性の高いデータやアルゴリズムの相互利用による新たな価値の創出や社会的課題の解決が促進されると考えています．今後は要素技術の研究開発やパートナーの皆様との検証評価をさらに加速させ，データ駆動型社会の１日も早い実現に貢献します．

（上段左から）大村　圭/ ジェイ　ホンジェ/ 片山　翔子/ 河井　彩公子
（下段左から）柏木　啓一郎/ 馬越　健治/ 除補　由紀子/ 木村　達郎

私たちは，本稿で紹介した次世代データハブの技術確立とその実用化を通じて，企業や組織の枠を越えたデータの流通と利活用による新たな価値の創出や社会的課題の解決に貢献していきます．

◆問い合わせ先
NTTソフトウェアイノベーションセンタ
企画担当
TEL　0422-59-2207
FAX　0422-59-2072
E-mail　sic＠hco.ntt.co.jp