Informatica

ホワイトペーパー

AI 技術によりデータ駆動型ビジネスを支援

CLAIRE の機械学習技術によりデータの生産性を飛躍的に向上させる方法

CLAIRE™

2017 年 5 月発行

目次

はじめに

デジタル変革（トランスフォーメーション）は、絵空事ではありません。すでに現実のものとなっています。もはや、問題となるのはこのような破壊的変革を「する側になるのか、される側になるのか」ということです。組織は変革のためのイニシアチブを推進して、財務業績や業界での競争力を高めています。企業が推進する変革のイニシアチブには、例えば顧客との関係の強化や業務の最適化、医療サービスのパーソナライズ、不正行為の防止などがあります。

これらのイニシアチブを成功へと導くためには、信頼できるデータをタイムリーに活用できる必要があります。すなわち、デジタル戦略の成功基盤となるのは、データです。デジタル戦略の成否は、データをどれだけ効果的に管理できるかによって決まります。言い換えれば、デジタル戦略の効果は、戦略基盤として情報をもたらすデータの効果に比例するということです。

ただし、「これまでと同じやり方」でデータを管理していても、成功は見込めません。IT リーダーは、データ管理の生産性を高め、より質の高いデータを、より迅速に、すべてのユーザーに提供できる方法を探しています。

インフォマティカの CLAIRE™（CLoud-scale AI-powered REal-time）エンジンは、人工知能（AI）と機械学習技術によって全社のデータとメタデータを活用することで、すべてのデータ管理者とデータユーザーの生産性を飛躍的に高めます。

データ管理のトレンド

データやデータアーキテクチャに対する考え方を見直す時がきています。これまで数十年にわたり、重点が置かれてきたのはビジネスシステムとプロセスでした。これらの重要性は依然として変わりませんが、市場で真の差別化を図るには、品質、適時性、完全性に優れたデータに基づいてビジネスイニシアチブを推進しなければなりません。しかし、ほとんどの場合 IT 予算の増加ペースは鈍いため、手元にあるリソースでより多くの成果を挙げる方法も検討する必要があります。

現在、企業データ管理はかつてないほど困難になっています。データの価値を最大限に引き出すためには、IT 部門は以下を管理できなければなりません。

1. **より多くのデータ**：
   - **データ量**：グローバル規模のデータセンターが処理するトラフィックは年間 15.3ZB（ゼタバイト）。
   - **データの複雑性と多様性**：企業の内外を通じて、さまざまな新しいデータソースとデータタイプが出現しています。
   - **データ速度**：モノのインターネット（IoT）の普及を背景に、常時オンラインの 200 億を数えるコネクテッドデバイスからのデータストリーミングをサポートすることが必要です。
2. **より多くのユーザー**：業務データユーザー数は 3 億 2,500 万人に上り、その数は増え続けています。業務アナリストや一般のデータサイエンティストからデータスチュワード（データ管理／案内人）まで、さまざまなユーザーがデータへの直接かつタイムリーなアクセスを必要としています。
3. **より多くの統合パターン**：
   - **クラウドへの移行**：ERP スイートは個々のソリューションに分けられ、クラウドに移行しています。
   - **アナリティクステクノロジー**：業界全体が、ビッグデータや NoSQL、予測アナリティクスなどの新しいテクノロジーに移行して、データウェアハウジングをサポートしています。
   - **実験**：ユーザーは、データを基盤に迅速に仮説を立て、試行錯誤を繰り返し、このプロセスを早いサイクルで繰り返したいと考えています。仮説に価値があるかどうかを証明する上で重要なのは、正確性よりもスピードです。

IT リーダーへの影響

デジタル変革の成否を左右するのはデータであると企業は認識しています。しかし、前述のトレンドによってデータ管理はいっそう複雑になっています。

これは、データ主導のリーダーシップを発揮して自社の成功を支援する絶好の機会でもあります。例えば、どうすれば IT リーダーは、高コストとなる大人数の開発者集団に頼ることなく、より優れたデータをより迅速に業務部門に提供できるでしょうか。

限りある IT 予算でこれを達成する方法は、主に 3 つあります。

- データ管理タスク／プロジェクトの自動化を進めて効率性を高める
- 業務担当者によるセルフサービスを促進する
- コラボレーションを増やして業務部門とIT部門の連携を強化する

ビジネスリーダーへの影響

ビジネスリーダーは、革新的なイニシアチブの推進が可能になり、これまではコスト面の理由からできなかった要求もできるようになったと感じています。しかし、基盤となるデータの質によってデジタルイニシアチブの結果は大きく変わります。

> **最優先すべきこと、それはすべてのデータの力を最大限に引き出す計画を策定することです。**

データ管理のコンピタンス（能力）を確立して、すべてのデジタルイニシアチブの基盤を固めることが重要です。データを資産として管理し、社内の誰でも探索したり利用したりできる環境を整える必要があります。そして、目的に応じたデータ品質（例えば、重要な意思決定やコミュニケーションには高品質のデータ／迅速なイノベーションや反復的作業には標準品質のデータ）を確保しなければなりません。また、技術面においては、手作業でのコーディングや統合されていないデータ管理ツールの寄せ集めでは、業務部門のニーズに対応できません。

機械学習とは

機械学習とは、プログラムが静的な状態にとどまることなく、データからの学習を何度も繰り返していく手法です。機械学習システムで構築した入力ベースのモデルは、予測や意思決定に利用できます。これらのシステムは、データから学習するだけでなく、自動的に調整してより良い結果を生み出すことができます。データの量が多いほど学習のスピードも上がり、結果の精度が高まります。

データ管理に機械学習が必要な理由

重要なビジネスイニシアチブに対してより迅速にデータを提供するには、自動化を進める必要があります。その際に役立つのが機械学習です。全社規模でメタデータへの可視性と機械学習があれば、データ管理ツールに「学習」させて、インテリジェントな提案を実行したり、さまざまなデータ管理タスクを自動化したりできるようになります。ただし、機械学習によってデータアナリストなどが不要になるわけではありません。機械学習はあくまでもデータ管理担当者の生産性や有効性を高めるための手段に過ぎません。

機械学習を利用することで、人間には煩雑であったり不可能であるような作業も対応可能になります。以下のような例があります。

1. 検出と特定

- データ品質ルール、ビジネスエンティティの検出
- セマンティック検索、パターン認識、データ分類
- 異常の検出と通知

2. 予測的オペレーション

- データスパイクの処理
- 運用面の課題調査を優先付け
- 環境の変化に対する自動対応力

3. 次にとるべき最善策と提案

- データセット、変換、ルールの提案
- ソースからターゲットへの自動マップ、クレンジング、標準化
- 新しいデータソースの自己統合

データ管理における機械学習の基盤

機械学習を効果的に行うためには、大規模なトレーニングデータセットが欠かせません。データ管理にとって最適なデータソース、それは全社レベルのデータカタログです。通常、企業には数千ものデータベース、データファイル、アプリケーション、アナリティクスシステムがありますが、これらのデータリポジトリにあるメタデータを収集することで、充実したカタログを作成できます。機械学習、データカタログ、また全社レベルでのメタデータへの可視性を組み合わせることで、インテリジェンスの基盤を構築して、データ管理の生産性を確実に高められます。

このアプローチは SaaS アプリケーションにも通用するので現在のクラウド時代においても有効で、Salesforce や Workday などの SaaS アプリケーションからメタデータを収集してエンタープライズカタログに追加できます。

Informatica CLAIRE：Intelligent Data Platform の「知能（=インテリジェンス）」

インフォマティカのアプローチでは、以下のように機械学習によってデータ管理の生産性を高めます。

1. Intelligent Data Platform（IDP）：インフォマティカは、生産性を最大限に高めるエンドツーエンドの統合データ管理プラットフォームを提供しています。接続性、メタデータ、運用管理を統合したプラットフォームによって、新しいデータ管理プロジェクトも迅速に開発して導入できます。このプラットフォームは、オンプレミス、クラウド、ビッグデータの各ソースのデータを管理する強力で一貫した機能を備えています。インフォマティカでは、この統合データ管理プラットフォームを「Intelligent Data Platform」と呼んでいます。

   このプラットフォームはモジュール式なので、任意の 1 つのツールから小規模に導入を始め、後は貴社のペースに合わせて徐々に拡張していくことができます。

図 1：データ管理機能、共有接続性、運用インサイト、データ／メタデータインテリジェンスを統合して提供する Intelligent Data Platform

2. メタデータ：インフォマティカは、長年にわたりテクニカルメタデータとビジネスメタデータの管理におけるリーダーとして広く知られていますが、この分野における能力をさらに高め、以下を含む広範なメタデータを全社レベルで収集することを可能にしています。

   - データベーステーブル、列情報、データプロファイル統計などのテクニカルメタデータ
   - データのコンテキストや意味、関連性、各種ビジネスプロセス／部門にとっての重大性を把握するためのビジネスメタデータ
   - システムやプロセスの実行に関する運用メタデータ（最終更新日時、ロードプロセスの最終実行日時、アクセス回数が最も多いデータなど）
   - ユーザーアクティビティに関する使用状況メタデータ（アクセスしたデータセット、クリックした検索結果、送信した評価やコメントなど）

このように広範なメタデータの収集は、機械学習に不可欠です。収集したメタデータを使用して機械学習のアルゴリズムが「学習」し、調整することで、より質の高い結果を生み出します。

3. インテリジェンス：CLAIRE は、メタデータと人工知能（AI）／機械学習を統合して提供します。

Intelligent Data Platform で収集したメタデータから得た広範な情報をベースに、CLAIRE のアルゴリズムが企業のデータ環境を学習します。この学習により、CLAIRE はインテリジェントな提案を行い、データ管理プロジェクトの開発と監視を自動化し、社内外の変化に適応します。Intelligent Data Platform のすべてのデータ管理機能のインテリジェンスを推進するのが、CLAIRE なのです。

CLAIRE の機能

CLAIRE は、幅広い範囲のユーザーに高い価値を提供します。

- データ開発者 ー 多くの実装タスクを部分的に自動化できるだけでなく、完全自動化することさえ可能です。
- データアナリスト ー 必要なデータをより簡単に見つけて準備することができます。
- 業務担当者 ー 規定のデータガバナンスおよびコンプライアンス管理の対象となるデータを迅速に特定できます。
- データサイエンティスト ー より短時間でデータへの理解を深めることができます。
- データスチュワード ー より簡単にデータの品質を可視化できます。
- データセキュリティ担当者 ー データの不正利用の検出と機密データの保護をより簡単に実現しながら、データが適切に管理されていることを周知できます。
- 管理者／運用担当者 ー データ管理プロセスのパフォーマンス最適化や予防保全を実行できます。

下記に、CLAIRE が実現するインテリジェンスの使用例をいくつか紹介します。

インテリジェントな類似データの提案機能

CLAIRE は、クラスタリングなどの機械学習技術を用いて、数千ものデータベースとファイルセットから類似データを検出します。インテリジェントな類似データの提案機能は、データの特定や重複の検出、また個別データ項目をビジネスエンティティにまとめる、タグを複数のデータセットに適用する、ユーザーにデータセットを提案するといった、さまざまな目的に活用できる主要な機能です。

類似データの提案機能では、2 つの列内のデータがどの程度類似しているかを自動計算します。企業環境で 2 列のペアをすべて比較する総当たり式のアプローチでは膨大な計算量となり（例えば 1 億列）、現実的ではありません。そのため、類似データの提案機能は、機械学習技術を使うことで類似する列をクラスタリングし、一致度の可能性が高い列を特定します。

このプロセスは、複数のステップで実行されます。まず、列を特徴に基づいてクラスタリングします。次に、データの重複を検証して、各クラスタ内に一意値があるかどうかを確認します。最後に、最も可能性の高いペアを選択し、Bray-Curtis 係数と Jaccard 係数を使用してデータの類似度を計算します。

タグを使用したインテリジェントなドメイン検出

CLAIRE では、各列にセマンティックラベルを付けてデータ項目を分類できます。このセマンティックラベルを「データドメイン」と呼びます。

通常は、正規表現や参照テーブル、その他の手作業でコーディングされた複雑なロジックに基づきルールを評価して、セマンティックラベルを付けます。数千ものルールを定義して保守することは大きな作業負荷となります。

そのため CLAIRE では、タグの概念を使用してデータ項目の検出とラベル付けのプロセスを大幅に簡素化しています。未分類の列には、列の内容を示すシンプルなタグ（例えば「保険金支払日」）をユーザーが付ければ、システムがこれを学習し、このタグをすべての類似列に自動適用します。これはデータ管理における「顔認識」機能のようなもので、例えば Facebook の写真にタグを付けると同一人物を写したその他の数百万もの写真にタグが付けられるのと同じです。

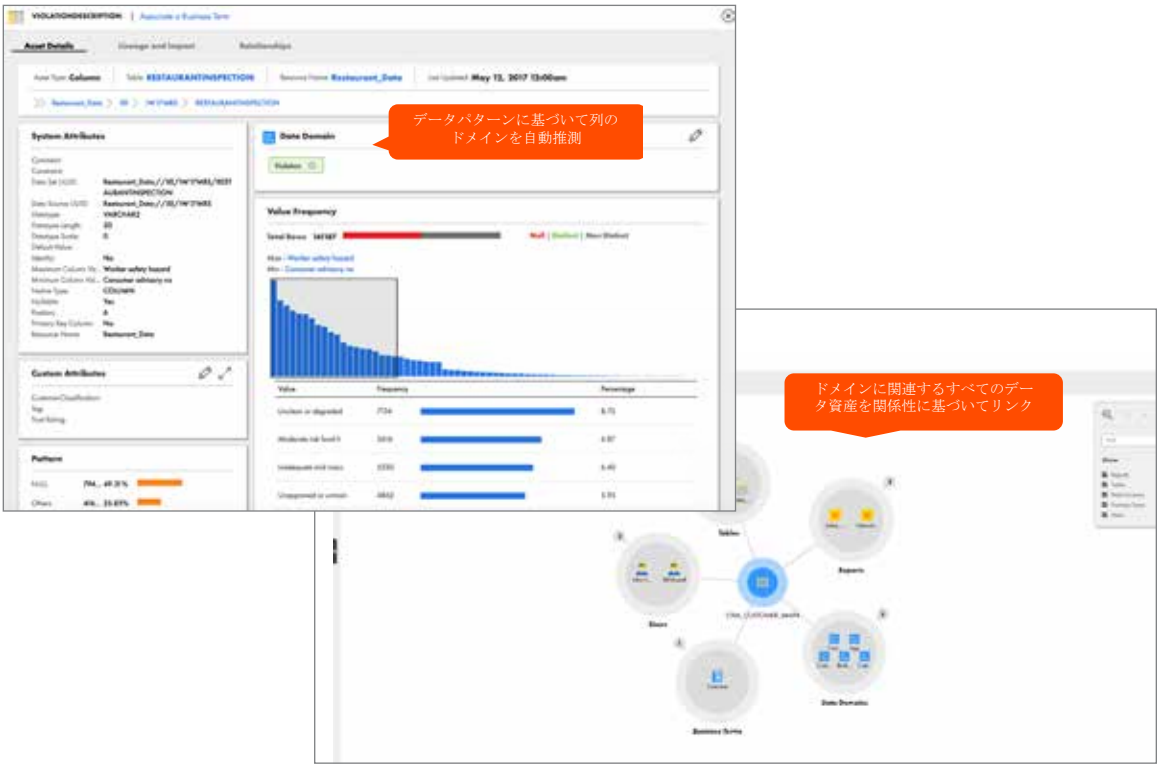

図 3：データの自動分類

インテリジェントなエンティティ検出機能

列のドメインを特定したら、CLAIRE で個々の項目を上位レベルのビジネスエンティティへとまとめることができます。下図（図 4）は、「顧客」項目と「製品」項目とを結合して「発注書」エンティティを作成する例です。エンティティ検出機能は、アナリティクスプロセスやデータ統合プロセスの際に、ユーザーがどのように異なるデータ項目をまとめたのかを学習し、その学習内容を適用して企業のデータ環境全体からエンティティを検出します。

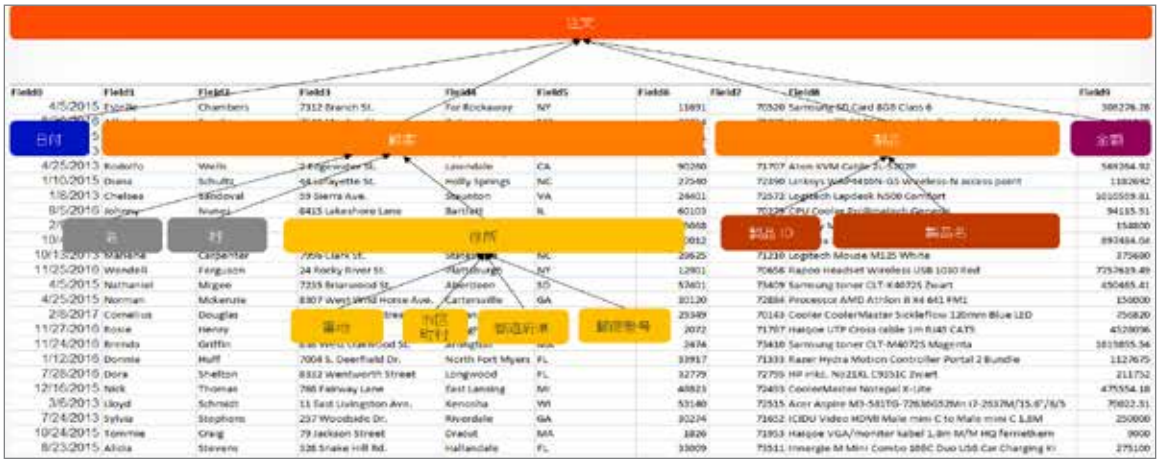

図 4：データドメインを結合してテーブルやファイルからエンティティを検出

インテリジェントなデータ提案機能

CLAIRE は、各プロジェクトで使用するデータセットをデータアナリストやデータサイエンティストに提案します。CLAIRE はユーザーが選んだデータセットを観察することで、より類似度が高いデータセットや補完となる追加データセットを提案します。このインテリジェントなデータ提案機能を使用すれば、社内で他のユーザーがすでに実行しているかも知れない作業を繰り返す必要がなくなります。CLAIRE の提案例は以下の通りです。

1. 同一データの準備済みバージョン（代替可能データ）

2. 同じタイプのレコードを有する別のテーブル（統合可能データ）

3. 属性を追加してデータをエンリッチ化するために結合するテーブル（結合可能データ）

データ提案機能は、コンテンツに基づくフィルタリングによって、追加のデータセットを提案します。データセットに適用される特性（条件）は、リネージ情報やユーザーによるランク付け、データ類似度などです。複数の類似度測定値によって異なるデータセット間の類似性がスコアリングされ、このスコアに基づいて類似するプロパティを有するデータセットが提案されます。また、メタデータグラフでクエリを実行することによって、複数のユーザーが共通して使用しているデータセットを見つけ出し、これらのデータセットを補完項目として提案します。

インテリジェントな構造検出機能

CLAIRE は、未整理のデバイスファイルやログファイルの構造を検出することで、ファイルを理解したり作業したりしやすくします。コンテンツをベースにしたアプローチでファイルを解析するため、ファイルの処理に影響を与えることなく、頻繁なファイル変更に対応できます。

インテリジェントな構造検出機能は、遺伝的アルゴリズムによってファイルのパターン認識を自動化します。このアプローチでは、「進化」の概念を使用して結果の質を高めます。ソリューション候補にはそれぞれ一連のプロパティがあり、このプロパティを変更およびテストすることで、ソリューションの適合性が高まるかどうかを判断できます。ファイルの構造を定義するのにユーザー入力は不要で、業界標準のファイル形式以外もサポートされています。ファイルの初期構造は、基本的なデリミター（区切り文字）ベースの解析に基づいて検出されます。検出された構造は、入力範囲や検出ドメインなど各種要素に基づいてスコアリングされます。最もスコアの高い構造が「変形」フェーズに移行して、構造にいくつかの変更が加えられます（サブ構造を組み合わせてスコアが増えるかどうかを試すなど）。データの適切な構造を特定できた段階でプロセスは終了となります。

図 5：非構造化データファイルの構造をインテリジェントに特定

インテリジェントな異常検出機能

CLAIRE は、統計アプローチと機械学習アプローチを用いてデータの異常値を検出します。ユーザー行動アナリティクス（UBA）機能により、リスクの高いユーザー行動やデータの不正利用につながる恐れのあるユーザー行動のパターンを検出できます。さらに、なりすましや認証情報のハイジャック、権限エスカレーション攻撃なども検出できます。

UBA 機能は、多次元モデルのユーザーアクティビティ（アクセスしたデータストアの数、実行したリクエストの数、各システムで影響を与えたレコードの数など）に対して、教師なし機械学習を実行します。このモデルに主成分分析（PCA）を実行することで、次元数を削減します。さらに、BIRCH 手法を用いた教師なしの階層的クラスタリングを通じて、指定の期間内に通常と異なる行動をしたユーザーを特定します。異常な挙動の検証には、距離と密度に基づく異常値検出法を用います。統計学に基づく Grubbs 検定を実施することで、前述の 2 つの手法で異常値と判断されたオブジェクトが実際にクラスタシステム内の異常値であることを確認します。

今後リリースが予定されている CLAIRE の機能のいくつかを下記に紹介します。

自己統合：新たに取り込んだデータを、データ統合プロセスへ自動的に統合します。数百万もの既存のマッピングとユーザー操作から学習して、データを特定し、類似データを処理する統合パターンを見つけ、データを自動的に変換して移動します。

開発支援：開発プロセスの際に、例えば下記のような次にとるべき最善策をユーザーに提案します。

- 変換の自動完了
- テンプレートの提案
- 機密データのマスキングタイプの提案
- クレンジングと標準化のためのデータ品質の提案
- パフォーマンスの自動最適化

自動マッピング：社内全体を通じてマスターデータエンティティを検出してマスターデータモデルへ自動的にマッピングし、必要な変換と品質ルールを適用します。

自己修復：メモリ不足や処理能力不足など、外部システムの問題を効果的に修復します。例えば、処理能力を追加（クラウドバースティング）して、データ量の急増に対応できます。

自己調整：履歴情報、現在のデータ量、使用可能なシステムリソースに基づきスケジュールやコンピューティングリソースを予測および調整して、パフォーマンス基準を満たします。

自己保護：機密データを自動的に検出し、保護されている領域から出る前にマスキングを行います。

インフォマティカについて

デジタル変革（トランスフォーメーション）は、世界を変えつつあります。エンタープライズクラウドデータ管理のリーダーであるインフォマティカは、時代をインテリジェントにリードする企業を万全の態勢でサポートすると共に、俊敏性を高め、新たな成長機会を実現するだけでなく、新たなモノを生み出すことさえ可能にする将来への洞察力を提供します。インフォマティカのソリューションを存分に活用してデータの価値を最大限に引き出し、次のインテリジェントな破壊的変革を推進してください。一度だけではなく、繰り返し何度でも。

結論

現在のデータ中心型のビジネス戦略は、データを基盤に構築されています。このため、競争優位性を得るには、データの力を最大限に引き出すためのデータ管理能力が不可欠です。

通常の環境において、データ管理ではさまざまな課題が発生しますが、従来型のアプローチは、将来は言うに及ばず現在の要件にも応えられるだけの拡張性を備えていません。このような中、データを活用して変革を進めるための方法の 1 つ、それがエンドツーエンドのデータ管理プラットフォームを基盤に標準化を実現することです。このようなプラットフォームでデータやメタデータ、機械学習／AI を活用することによって、すべてのユーザー（技術、運用、業務）の生産性を高められると共に、特に業務部門のユーザーのセルフサービスを促進することが可能になります。

CLAIREとIntelligent Data Platform によって、どのようにデータの価値を最大限に引き出して活用するのか。詳細は、お問い合わせください。

〒105-6226 東京都港区愛宕2-5-1 愛宕グリーンヒルズMORIタワー26階 電話：03-6403-7600(代表) FAX：03-3433-1021
www.informatica.com/jp linkedin.com/company/informatica twitter.com/Informatica

IN09_0517_3328