Informatica

ホワイトペーパー

人工知能が組織をデータドリブンでインテリジェントに変える方法

CLAIREのAI／機械学習による革新がデータマネジメントの進化を推進する

インフォマティカについて

デジタルトランスフォーメーションによって我々の期待値が変化しています。より良いサービスを、素早く、便利に、低コストで利用したいという期待が高まっているのです。企業も状況に応じて変化する必要があります。そしてそのヒントは「データ」にあります。

エンタープライズクラウドデータマネジメントで世界をリードするインフォマティカは、俊敏性の向上、新たな成長機会の獲得、新しいソリューションの開発を実現するための洞察を通じて、あらゆる産業や分野の企業がインテリジェントにビジネスをリードできるよう支援します。インフォマティカは、あらゆるデータを徹底的に重視し、企業の成功に必要とされる汎用性を提供します。

インフォマティカは、企業がこれからのインテリジェントな破壊的イノベーションを推進できるよう、当社が提供するあらゆるサービスを通じてデータの力を継続的に引き出すことを支援します。

目次

AIの重要性 4
AIにはデータが不可欠 4
データにはAIが不可欠 5
Informatica CLAIRE：Intelligent Data Management Cloudの
「知能（＝インテリジェンス）」 8
CLAIREを活用したデータのカタログ化 9
CLAIREを活用したアナリティクス 13
CLAIREを活用したマスターデータマネジメント 17
CLAIREを活用したデータガバナンスおよびコンプライアンス 19
CLAIREを活用したデータプライバシーとデータ保護 23
CLAIREを活用したDataOps 27
CLAIREの未来 28
結論 29

「データおよびアナリティクスのリーダーは、データ環境の複雑さに直面しています。当社は、これからのデータマネジメントソリューションに求められる機能として、AIを活用した定型業務／非定型業務の自動化、クラウド機能、接続されたデータアーキテクチャ、メタデータを予測しています」[1]

– Gartner社

AIの重要性

現在、世界中のあらゆる業界でデジタルトランスフォーメーションが推進されています。そして、その原動力となっているのが、人工知能（AI）と機械学習（ML）です。経営幹部がビジネス変革戦略において最も重視しているのがAIです。映画鑑賞から自動車の運転まで、AIは、私たちの日常生活に広く浸透しており、より豊かな生活をもたらしています。AI/MLは、ライフサイエンス分野で新しい治療法を発見したり、金融サービスで詐欺やリスクを軽減したり、高度にパーソナライズした顧客体験を提供したりする上で不可欠な存在となっています。

AI/MLは、ビジネスリーダーの目には魔法のように映るかもしれません。大きな可能性を秘めていることは明らかだが、その正体を十分には把握できない、または効果的に使いこなす方法が分からないと感じているリーダーもいることでしょう。AI/MLは、次にとるべき最善策の提案、顧客満足度の追跡、業務の効率化、革新的な製品の開発などを可能にする、各種新規ビジネスソリューションの基盤となるテクノロジーです。一般に、機械学習（特に深層学習）は大量のデータを必要とします。求められる精度を実現するためには、膨大な量のデータでMLをトレーニングする必要があります。そして、トレーニングに使用するデータは、ビジネスの現状を正確に反映したものでなければなりません。AIのトレーニングに粗悪なデータや不十分なデータを使用すると、ビジネスイニシアチブに悪影響が生じ、目的とする成果を達成できないどころか逆効果になる場合もあります。

効果的なAIを構築するためには、組織内外の多様なデータにアクセスして、適切な特徴を使用し、適切なトレーニングを行う必要があります。そして、取り込んだデータに基づいて、MLモデルを構築およびトレーニングしなければなりません。そのために必要なのがデータ管理です。ここで重要なのはデータの拡張性と複雑性だけではありません。データの信頼性も重要です。モデルのトレーニングに使用するデータは、適正なシステムから取得したデータか？個人を特定できる情報（PII）を取り除いて、すべての規制に準拠しているか？透明性を確保しているか？モデルに使用するデータのリネージを証明できるか？データにバイアスがないことを示すための文書を作成したか？その文書を規制当局や調査当局に提出できるか？これらを実現するためには、適切な制御とデータ管理のための基盤が必要です。強固なデータ管理基盤がなければ、AIを理解することも信頼することもできません。つまり、データ管理がなければ、AIはブラックボックスと化し、予期せぬ結果をもたらす可能性があります。

AIにはデータが不可欠

AIの成否は、データサイエンティストがどれだけ効果的なモデル（AIのトレーニングおよび拡張のためのモデル）を設計できるかによって決まります。そして、モデルの成否は、どれだけ信頼できるデータをタイムリーに取得できるかによって決まります。

データサイエンティストがAI/MLモデルを構築する際に高品質データを必要とするのはなぜなのでしょうか？消費者の行動を予測するための予測モデルを例に考えてみましょう。このような予測モデルにとって、例えば郵便番号で示す消費者の所在地は有用な特徴となります。しかし、その郵便番号データが不明であったり不完全または不正確であるとどうなるでしょう？予測モデルの挙動（トレーニング時および実装時の両方）に悪影響が生じ、不正確な予測につながり、イニシアチブ全体の価値が低くなってしまいます。正確、完全、検証済みの郵便番号データがあれば、消費者の市場セグメント、所得クラス、年齢、平均余命などを予測できます。今後、「説明可能なAI」は、任意のオプションではなく、必須になるはずです。メタデータ主導のリネージとトレーサビリティがなければ、AIを活用したアプリケーションとインサイトを本番環境に実装することはできません。

[1] Gartner社『Predicts 2020: Data Management Solutions』Rick Greenwald氏、Donald Feinberg氏、Mark Beyer氏、Adam Ronthal氏、Melody Chien氏共著、2019年12月5日

AIにはインテリジェントなデータ管理が必要です。インテリジェントなデータ管理により、モデルに必要なすべての特徴を迅速に見つけ、AIモデルのニーズ（特徴の拡張、標準化など）に合わせてデータを自動的に変換し、データの重複を排除して信頼できるマスターデータ（顧客データ、患者データ、パートナーデータ、製品データ）を提供し、（モデル内データおよびモデルの運用に必要なデータの）エンドツーエンドのデータリネージを提示する必要があります。AIの成否は、データサイエンティストがどれだけ効果的なモデル（AIのトレーニングおよび拡張のためのモデル）を設計できるかによって決まります。そして、モデルの成否は、どれだけ信頼できるデータをタイムリーに取得できるかによって決まります。

データにはAIが不可欠

AI/MLは、データ管理を拡張する際にも重要な役割を果たします。デジタルトランスフォーメーションでは膨大な量のデータが必要になるため、最も関連性の高いデータとメタデータを検出およびカタログ化して、その関連性、価値、セキュリティを証明し、透明性を確保する必要があります。さらに、このデータをクレンジングしてマスター化し、効果的にガバナンスおよび保護しなければなりません。データ（およびデータの拡張）を効果的に管理しなければ、AI/MLモデルは従来のデータウェアハウジングイニシアチブが過去30年間に苦しんできたのと同じ運命、つまり粗悪なデータを使用して、信頼性の低いインサイトを提供するという状況に陥ることになります。

最近の調査によると、データセンターの全体的なトラフィック量は2021年までに20.6ゼタバイト（ZB）に達し、コネクテッドデバイスと接続の数は2022年までに250億に達すると予測されています[2]。ガバナンスポリシーに準拠しながら、これらすべてのデータを処理して、使用可能／信頼可能な形に変換しなければなりません。加えて、俊敏性を維持して、ビジネス戦略やビジネスプロセスの変化にも対応する必要があります。データの増加に伴い、デジタルトランスフォーメーションイニシアチブのためのデータ準備作業はますます複雑化しています。LinkedIn社によると、米国で最も有望視されている職種のひとつがデータサイエンティストです。[3]また、企業が必要としているデータエンジニアの数は前年比で96%増加しています。[4]しかし、これらの人材を採用しただけでは、増加するデータ量を管理することはできません。

直線的なアプローチでは指数関数的課題には対応できない

このような課題は、単純にエンジニアや開発者の数を増やせば解決できるというものではありません。従来型のアプローチは非効率性に満ちています。プロジェクトはサイロ化されており、メタデータのエンドツーエンドの可視性はほとんどなく、自動化も限られています。機械学習もなく、処理は高コストで、ガバナンスとプライバシーの手順は繰り返し何度も実行しなければなりません。では、ビジネスのスピードに合わせて、セルフサービス、顧客サービスの強化、業務の生産性の向上、迅速なイノベーションを実現するにはどうすればいいでしょうか？

[2] Cisco社『Global Cloud Index Forecast and Complete Visual Networking Index Forecast』
[3] LinkedIn社『LinkedIn’s Most Promising Jobs of 2019.』
[4] Datanami社『Data Engineering Continues to Move the Employment Needle.』

ここで威力を発揮するのがAIです。AIにより、データ管理に関するタスク（データの検出、統合、クレンジング、ガバナンス、マスタリング）を自動化および簡素化できます。単調で反復的なタスクを機械学習に学習させて代わりに実行させることで、開発者とユーザーはより価値の高い革新的なプロジェクトに専念できます。AIにより、データに対する理解を深め、データのプライバシーや品質に関する異常を特定できます。AIは、開発者、アナリスト、スチュワード、業務担当者の最適なパートナーとなり、自動化や強化（提案、次にとるべき最善策）を通じてタスクの高速化に貢献します。

AIを最大限活用するためには、どうすればAIを使用してデータ環境全体でエンドツーエンドのプロセスを高速化できるかを考える必要があります。AIはデータ管理に不可欠です。このため、インフォマティカはCLAIRE®エンジン（インフォマティカのメタデータ主導のAI機能）のイノベーションに重点的に取り組んできました。CLAIREは、全社レベルの統合メタデータを活用して、日常的なデータ管理／スチュワードシップタスクの自動化と拡張を行います。

AIを活用したデータ管理の4つの主なメリット

一般に、データ管理チームはAIによって「データ担当者の生産性を高める」、「業務効率を向上させる」、「データ体験およびデータに対する理解をインテリジェントに改善する」、「データガバナンスのプロセスを加速させる」という4つのメリットを享受できます。AIによって可能になることのいくつかの例を以下に紹介します。

生産性：データ統合の提案システムにより、データエンジニアはデータを抽出、変換、提供するためのマッピングを迅速に構築できます。提案システムが既存のマッピングを学習し、データベースやファイルシステムのビジネスコンテンツを理解し、ターゲットとなるシステムやデータ利用者に配信する前に、データの標準化やクレンジングのための適切な変換を提案します。

効率性：一般的な企業では、数千ものデータ統合プロセスが日々実行されています。これらのプロセスのモニタリングは、多くの場合、管理ツールが所要時間とCPU／メモリ使用率を記録するだけの受動的な作業です。AIがログファイルやモニタリングファイルの時系列データの履歴値を学習し、異常値にプロアクティブにフラグ付けし、事前に対処されない場合に起こり得る問題を予測します。

データ体験：実際のエンティティ（患者記録や受注記録など）をデータベースやファイルに保存すると、エンティティ内のデータは断片化され、複数のテーブルやファイルに分散して保存されます。これは、ストレージとパフォーマンスを最適化するためです。AIにより、これらの分散したデータの関係を検出し、元のエンティティを迅速に再構成できます。ユーザーは、プライマリキーと外部キーの関係に関する古い文書を思い出したり、探したりする必要はありません。また、データセットを手作業で結合する必要もありません。さらに、AIは、類似のデータセットを特定して、使用パターンやデータ品質、クラウドソーシングしたコラボレーションに基づいて提案を行います。

データガバナンス：データガバナンスでは、ユーザーがデータを理解できるように、ビジネス用語と物理データ要素を関連付けて、データ要素のビジネスコンテキストと関連性を確立するという単調な作業が発生します。多くの場合、AIにより、複数の自然言語処理（NLP）手法とビジネスタイプ特定を用いて、ビジネス用語と物理データの関連付けを自動化できます。これにより、この単調でミスの発生しやすい作業を大幅に軽減できます。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。

AI主導のデータ管理：銀行業界の例

なぜAIにはデータ管理が不可欠なのでしょうか？ある銀行を例に見てみましょう。

より多くのデータにAIを適用して、高度、予測的、リアルタイムのアナリティクスを実行することにより、銀行では次のことが可能になります。

- よりパーソナライズしたサービスを提供して、顧客維持率を高める
- 販売時点における不正トランザクションを抑制する
- 資産運用アドバイザーのコストを削減しながら、顧客の投資結果を改善する
- プロジェクトに関連する法規制へのコンプライアンスコストを削減する

AIを活用したデータ管理では、あらゆるタイプの関連データ（ERP、CRM、クラウド／Webアプリ、マシン／ログファイル、サードパーティのデータなど）を自動的に検出およびカタログ化できます。これにより、データサイエンティストは必要なデータへ迅速にアクセスして、多数の実験を行い、パターンを特定し、顧客行動や不正活動、顧客のリスク傾向に応じた投資機会などに関するインサイトを獲得できます。

データ管理にAIを活用することで、顧客データ間の関係を探索して、インサイトを特定の人物に結び付けて、顧客や犯罪に関わる可能性がある人物の360度ビューを自動的にエンリッチ化できます。その結果、より関連性の高いサービスを提案して、各種チャネル（オンライン、モバイル、電話など）でシームレスな体験を提供し、顧客エンゲージメントを強化できます。犯罪に関わる可能性がある人物の360度ビューにより、銀行は不正活動のパターンやネットワークをより迅速に発見して、数百万ドル規模の損失を回避できます。

また、AIにより、データ統合／データ品質タスクを自動化および簡素化して、数百ものデータソースから収集したデータを結合およびクレンジングし、アナリティクスモデルおよびアルゴリズムの予測力を強化できます。より豊富かつ高品質のデータに加え、AI/MLと高度なアナリティクスを活用することにより、次の最適な提案や不正活動の特定などにおいて、優れた結果を達成できることが実証されています。

また、AIにより、データガバナンスを強化して、ポリシーを確実に文書化するだけでなく、確実に適用できます。これにより、情報セキュリティ担当者は、EU一般データ保護規則（GDPR）、米国企業改革法（SOX）、Basel IIおよびBasel IIIなど、データプライバシーに関する規制を遵守できます。

Informatica CLAIRE : Intelligent Data Management Cloud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 Management Cloud™ : 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かつクラウドネイティブの統合データマネジメント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、マルチクラウド、マルチハイブリッド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、この統合データマネジメントプラットフォームを「Intelligent Data Management Cloud」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開始し、後はそれぞれ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。

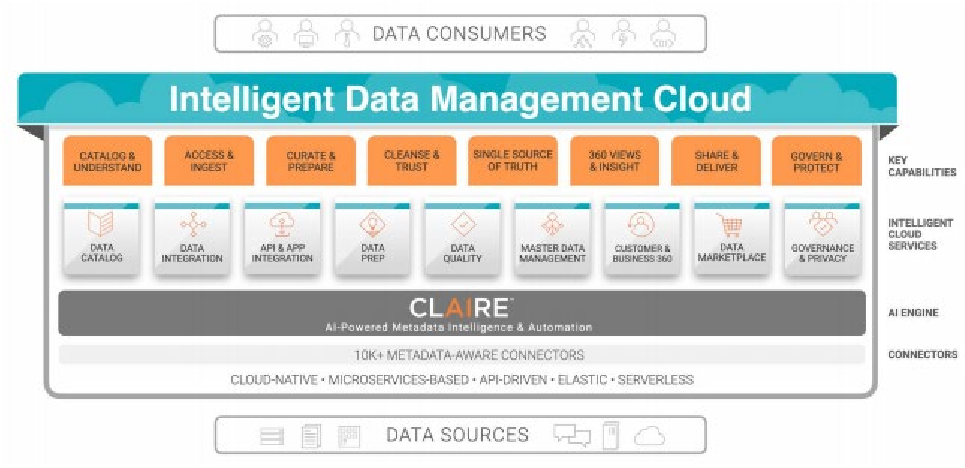

図1 : データ管理機能、共有接続性、運用インサイト、データ／メタデータインテリジェンスを統合して提供するIntelligent Data Management Cloud

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、スクリプト、データリネージなどのテクニカルメタデータ
   - 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など、データのコンテキスト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（データの最終更新日時、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 Management Cloud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モニタリングを自動化し、社内外の変化に適応します。Intelligent Data Management Cloud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。
- データエンジニア ― 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 ― 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 ― 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 ― 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 ― 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／プライバシー担当者 ― 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 ― 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。

ここで、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。

CLAIREを活用したデータのカタログ化

あらゆるデータイニシアチブにおいて、保有するデータを探索して理解することが最初の一歩となります。CLAIREにより、機械学習ベースの探索エンジンを使用して、社内のデータ資産をスキャンしてカタログ化できます。CLAIREを搭載したインテリジェントなデータカタログにより、データサイエンティスト、アナリスト、データエンジニアは必要なデータを検索して提案できるため、データの探索と準備に費やす時間を大幅に短縮できます。

高度な関係探索

データをカタログ化およびモデル化する際の重要なタスクとして、データセット間の関係を文書化する作業があります。CLAIREは、機械学習技術を用いて、構造化されたデータセット間のプライマリキー、ユニークキー、結合を自動的に特定し、数か月を要していた文書化を数分レベルに短縮します。CLAIREは、人間をデータキュレーションプロセスに関与させることで、関係を特定する能力を継続的に強化していきます。例えば、CLAIREが推測した関係をユーザーが承認または却下すると、CLAIREがそれを学習します。

銀行のデータアナリストが、マーケティングキャンペーンに反応する可能性が最も高い顧客に関するレポートを作成する際には、全顧客の購入済み商品／融資情報が必要です。しかし、これらのデータが複数の部門やデータストアでサイロ化していると、検索するのが困難です。CLAIREは、データベース内で文書化された結合、BIやETLなどのツールで実行された結合、またデータ値から抽出した統計情報に基づいて、結合を推測し、データアナリストに提案します。これにより、ユーザーの分析を拡大して、利用可能なすべての情報を活用し、キャンペーンの適正なターゲットを特定できます。

CLAIREは複数の手法を用いてキーと結合を探索します。キーについては、一意性、nullカウント、列のメタデータ（例：「ID」を含む列名）などのようなプロファイリング統計情報を組み合わせることで、プライマリキーとユニークキーを探索します。その後、結合および結合キーを推測する際に、列の署名分析などの機械学習技術を用いて、多数の潜在的なデータセットにわたって結合を探索します。

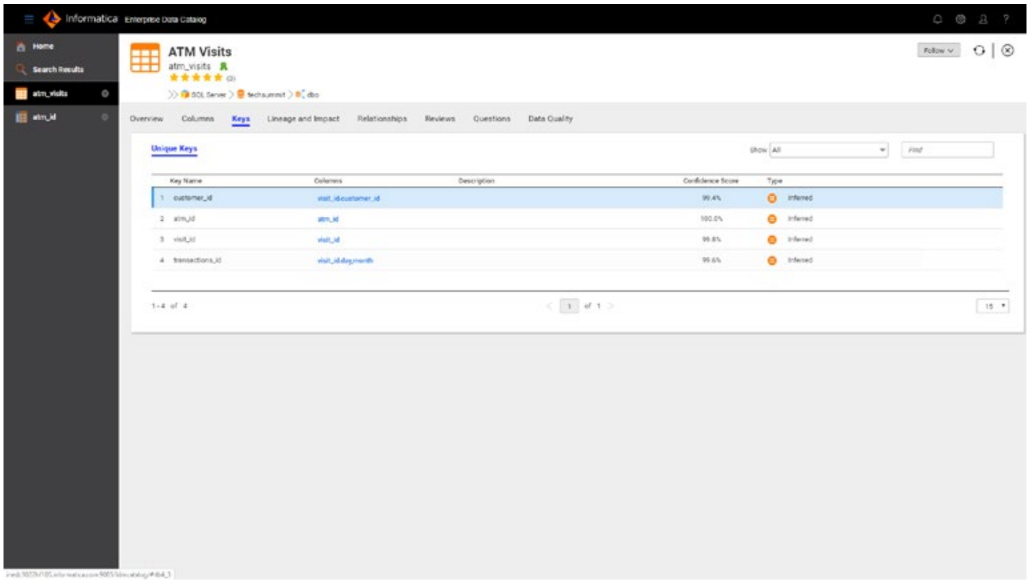

図2：機械学習技術を用いた推測を通じて、ユニークキーを探索

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。

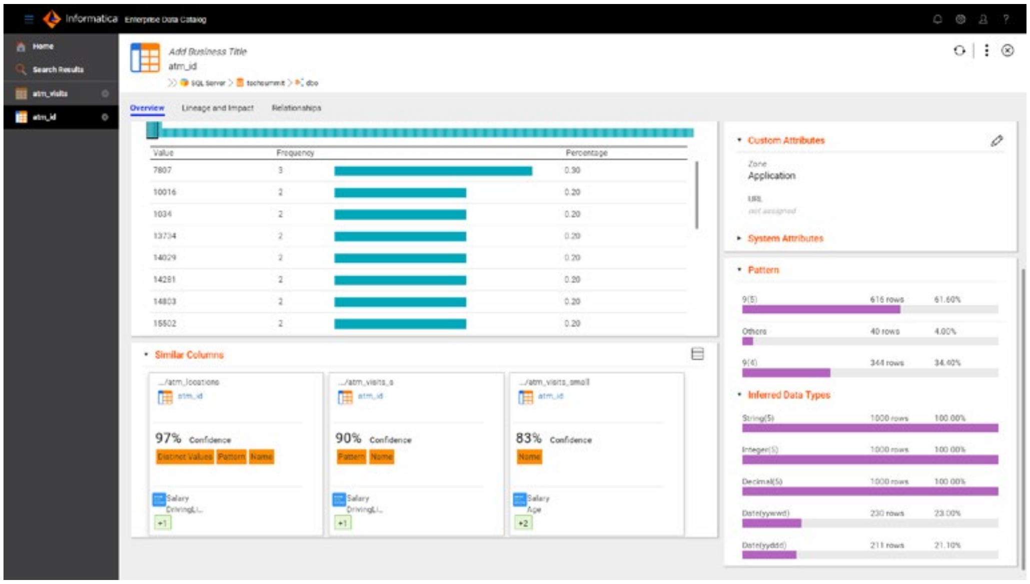

図3 : クラスタリング、Bray-Curtis係数、Jaccard係数を使用して、類似する列を特定

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。

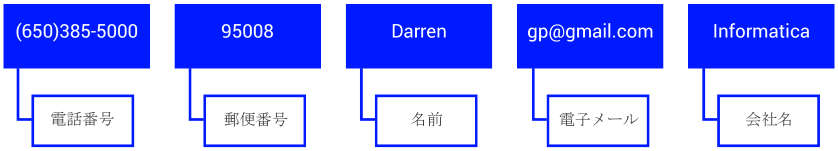

図4 : CLAIREは自動的にデータ項目を分類して、「タグ」と呼ばれるセマンティックラベルを適用

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。

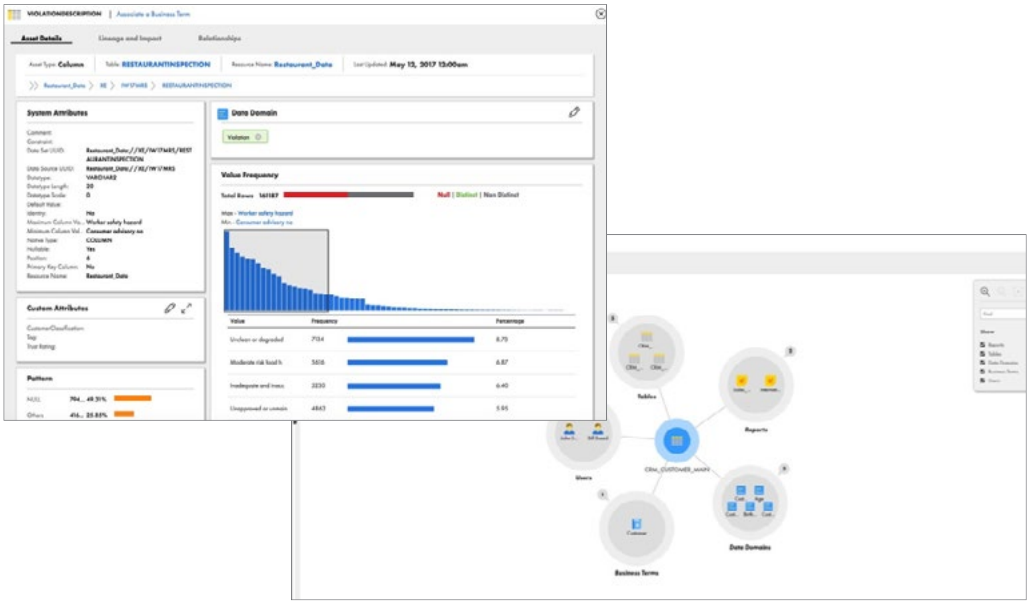

図5 : データの自動分類

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6 : 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

CLAIREを活用したアナリティクス

CLAIREを活用した自動化とインテリジェンスにより、アナリティクスのインサイトとプロセスを大幅に高速化し、データの可用性を高めることができます。また、アナリティクスのためのデータ準備も整流化できます。さらに、データパイプラインを提案する機能や複雑なマルチ構造化データを自動的に解析する機能により、データエンジニアリングの生産性を強化できます。

トランスフォーメーションの提案

データ統合マッピングの自動作成機能と次のトランスフォーメーションや式の予測機能により、設計のループを閉じて、データエンジニアの生産性を強化できます。CLAIREベースの提案機能をオンにすると、自社データパイプライン内の匿名のメタデータが分析され、AI/MLに基づいて設計に関する提案を受け取ることができます。このメタデータに基づいてトランスフォーメーションや式に関する提案が生成されます。提案の承認や却下をCLAIREが学習するため、提案機能を使用するごとにCLAIREの能力が強化されます。これにより、開発を高速化して、反復的なタスクを自動化し、さまざまなタイプのユーザーが迅速にデータを接続および統合できるようになります。

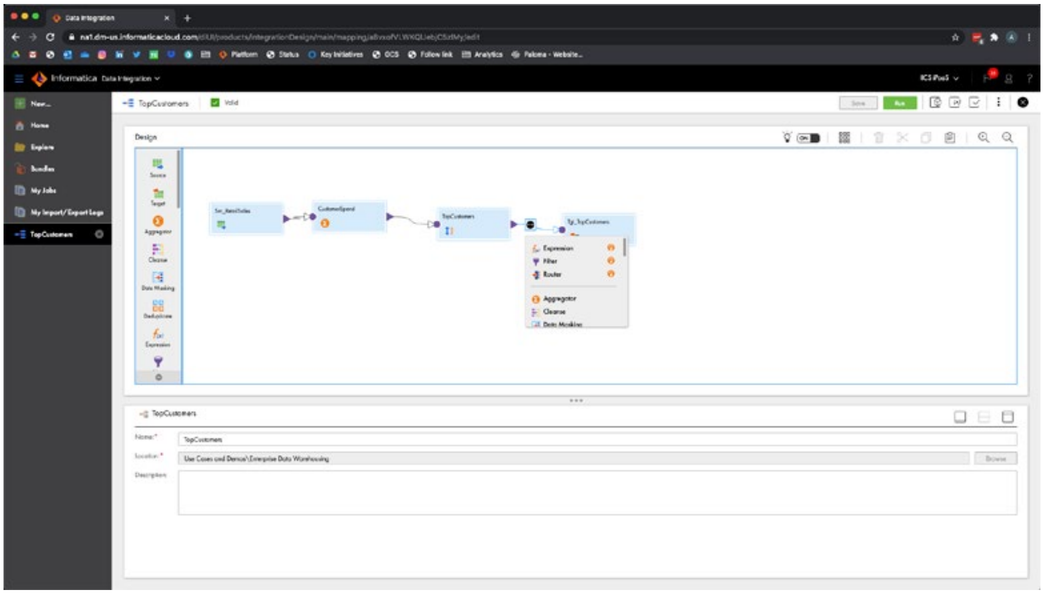

図7：データパイプラインの作成時にCLAIREが次の最適なトランスフォーメーションを提案

規模に応じて最適化されたプロセス実行

CLAIREは、さまざまな最適化手法を活用して、データパイプライン全体の統合パフォーマンスを強化します。スマートな最適化機能が、パフォーマンス特性に基づいてビッグデータワークロードを実行するのに最適な処理エンジンを判断し、過去のユーザーアクティビティに基づいてマッピングに関する提案をデータエンジニアに提示します。また、コストベースの最適化機能とヒューリスティクスが、データパイプライン内の結合の順序をインテリジェントに変更し、パフォーマンスを最適化します。これらは、データパイプラインを最適化するCLAIREの機能の一例に過ぎません。

結合列の提案

ユーザーが2つのデータセットを組み合わせるアクションを実行すると、CLAIREが結合列（結合キー）を自動的に提案します。これにより、データアナリストはデータセットを複合データセットにマージするための最適な方法を決定するのに数百時間も費やす必要がなくなります。CLAIREはまず、データレイクにインポートされたデータセットの元のソースシステム（例：Oracleなどのリレーショナルデータベース）で定義されたプライマリキーと外部キーの関係（Pk-Fk）を特定します。他のプロジェクトで同じデータセットが結合されている場合は、その結合列情報を使用して提案を行います。CLAIREは、これらすべての情報の処理とランク付けを行い、2つのデータセットに最適な結合列を提案します。さらに、データセットのサンプリングに基づいて、提案した列のデータ重複率も表示します。

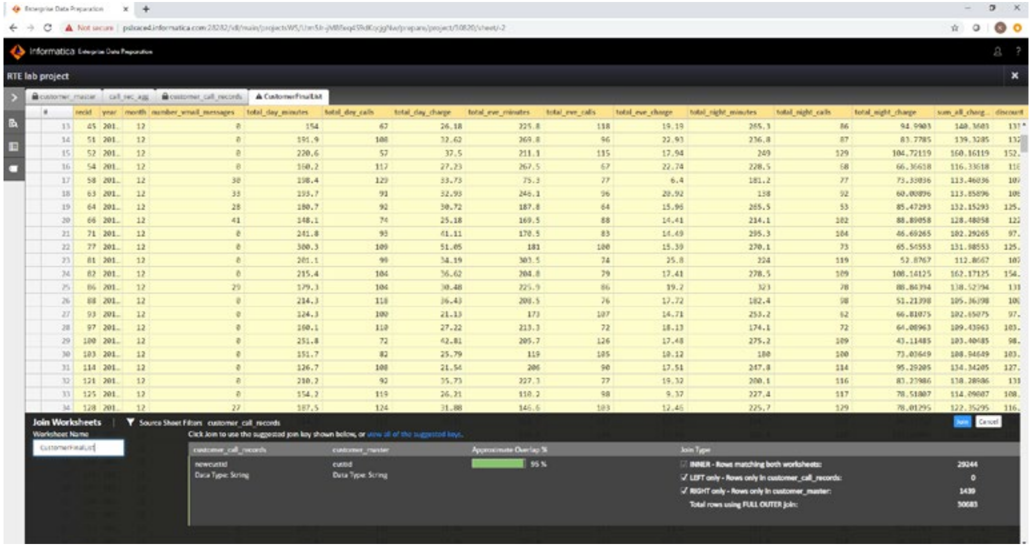

図8：2つのデータセットを組み合わせる際に、結合列を自動的に提案

Apache Zeppelinによる視覚化に関する提案

Informatica Enterprise Data Preparationは、Apache Zeppelinを使用して、グラフやチャートを含むノートブックの形式でワークシートを表示します。ユーザーがパブリケーションのノートブックを開くと、視覚化に関するCLAIREの提案が表示されます。パブリケーションのノートブックを初めて開く場合は、数値列のヒストグラムが表示されます。パブリケーションに数値列が含まれない場合は、ノートブックの最初のパラグラフに「SELECT * FROM」テーブルクエリが表示されます。

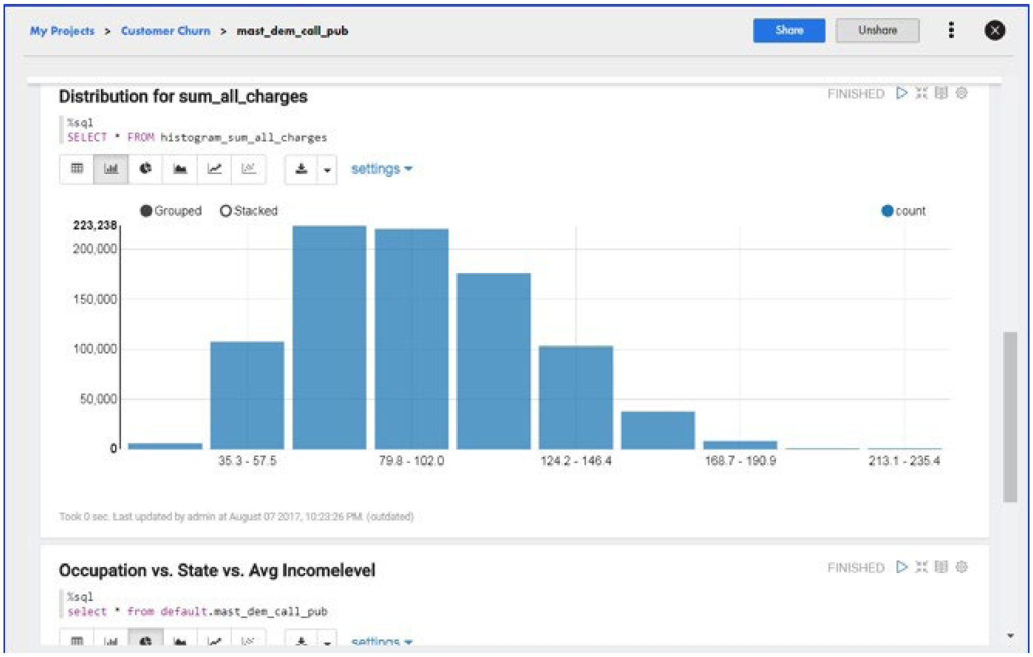

図9 : Apache Zeppelinノートブックによる視覚化に関する提案

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。

- 同一データの準備済みバージョン（代替可能データ）
- 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）
- 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。

Intelligent Structure Discovery

マシン、エンタープライズ、アプリケーションで生成および収集される非構造化形式または非リレーショナル形式のデータが増加しています。このようなデータタイプは、量が多いだけでなく、速度、多様性、変動性にも富んでいるのが特徴です。この新しいタイプのデータは、形式、速度、内容の変動が激しく、これを表す用語として「データドリフティング」という表現が一般的に使用されるようになっています。

CLAIRE搭載のInformatica Intelligent Structure Discovery（ISD）により、ファイルの取り込みとオンボーディングを自動化して、複雑なファイルを探索および解析できます。ISDは、クリックストリーム、IoTログ、CSV、テキスト区切り、XML、JSON、Excel、ORC、Parquet、Avro、PDFフォーム、Wordテーブルファイルなど、さまざまなデータファイル形式を標準でサポートしています。CLAIREによってこれらのファイルの構造が自動的に抽出されるため、簡単に把握して作業できます。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。

ISD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ます。このプロパティを自動的に変更し、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。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。

ISDは、カスタムMLベースの固有表現抽出（NER）と自然言語理解（NLU）メカニズムを利用して、項目と項目タイプを特定することで、その後の統合作業を簡素化します。また、社外アプリケーションもISDをNER/NLUエンジンとして利用できます。例えば、ISDを使用して、送受信するAPIペイロードに含まれるPIIを検出し、法規制へのコンプライアンスを確保したり、エンタープライズセキュリティを強化したりできます。また、ISDは、データの探索、取り込み、ストリーミングなどにも使用されます。

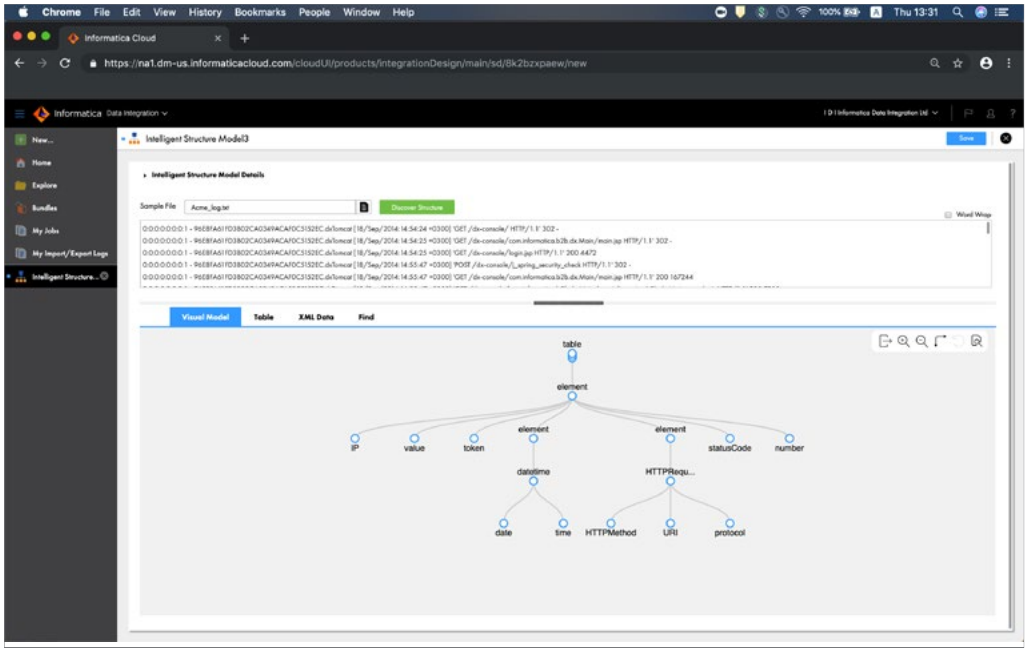

図10 : 非構造化データファイルの構造をインテリジェントに特定

CLAIREを活用したマスターデータマネジメント

CLAIREの高度なAIと機械学習を活用した自動化とインテリジェンスにより、顧客、製品、サプライヤーなどの分野の360度ビューをエンリッチ化して、精度を高めることができます。決定論的／発見的／確率論的アルゴリズム、コンテキスト統合照合、能動学習エンティティ照合など、さまざまなAI/ML技術が搭載されているため、高速、拡張可能、正確な方法を通じて、マスターデータのレコード照合とエンリッチ化を実行できます。

統合照合

統合照合とは、例えば見込み客を顧客に照合したり、インタラクションや非構造化データを顧客に照合するなどして、隠れた関係を発見するための次世代照合技術です。「コンテキスト属性」、機械学習、NLP、確率論的照合、宣言的ルールを活用します。

人口統計学的属性（名前、住所、電話番号など）、インタラクション属性（電子メール、チャットなど）、コンテキスト属性（いつ、何を、どこで、誰がなど）は、顧客関連データを高い信頼性で照合する上で大いに役立ちます。NLPにより、「コンテキスト属性」を非構造化テキストから抽出して、より多くのデータポイントを照合プロセスで使用できます。MLアルゴリズムは、適切に選択した照合ペアをデータスチュワードや各分野の専門家がラベル付け（一致または不一致）する教師あり学習を用いる場合に非常に効果的です。ラベル付けした照合ペアが、マッチングアルゴリズムの生成に使用されるトレーニングセットを形成します。

統合照合により、人口統計データ、アカウントデータ、トランザクションデータ、インタラクションデータ、非構造化データから成る360度の顧客ビューを構築できます。従来のマッチングアルゴリズムが複数のレコードをマージして、ひとつの顧客ビューを構築するのに対して、統合照合はすべての顧客データをグラフ形式で管理します。データは信頼度と共に関連付けられ、そこで顧客の複数のビュー（「視点」）を構築できます。

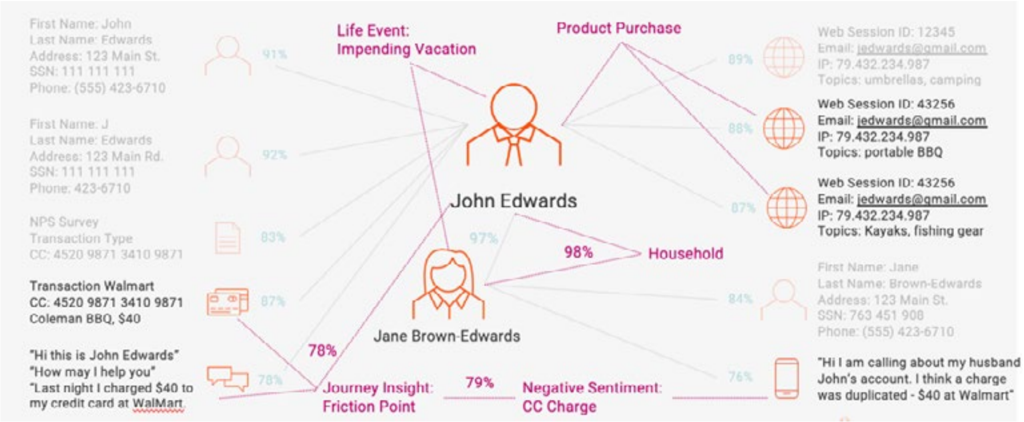

図11 : 統合照合および論拠によって推測したインテリジェンスを、Customer 360の一部として保存

ID照合

NAME3はCLAIREのID照合機能です。インデックス化とブロックのためのスマートなキー生成、パーティおよびロケーションデータのセマンティックテキストの安定化と比較、編集リストとテキスト安定化ルール（80人に対応）、特徴の重要度のインテリジェントな重み付け（目的別）など、さまざまな手法を用いて、30年以上にわたってトレーニングとチューニングを行っています。これらのパワフルな手法により、複数の項目のインデックス化とブロック、クライアント定義の一致／不一致ルールを指定した要件、およびその他のAIルールの補完となる実装時定義の一致／不一致ルールが可能になっています。

エンティティ照合

エンティティ照合で、同一のエンティティ（顧客、製品など）を指すデータレコードを検索できます。構造化データレコードと非構造化データレコード（チャット内に隠れた顧客情報など）の両方に対応しています。照合分類により、照合ペアを比較し、一致／一致の可能性／不一致を信頼度と共に判断します。人間が設定したルール（宣言的ルール）を使用する手法と、AIルール（機械学習による設定）を使用する手法があります。両方の手法を組み合わせることで、最良の照合結果が得られます。

各分野の専門家が作成した宣言的ルールは、CLAIREがランダムフォレスト分類として採用するパワフルなAIルールの補完となります。CLAIREは、クラウドソースまたはマルチユーザー学習ではなく、教師あり能動学習を使用して、AIトレーニングプロセスを促進します。10または20の照合ペアから成るマイクロバッチをユーザーにラベル付けします（一致／一致の可能性／不一致）。ラベル付けが完了すると、CLAIREはランダムフォレスト分類の再トレーニングを行い、次の最適な照合ペアを判断して、それをユーザーにラベル付けし、このラベル付けのプロセスを反復します。CLAIREは、このラベル付けされたペアを使用して、ブロックルールの推測（明らかな不一致の削除）、ブロックの実行、モデルのトレーニング、エンティティ照合の実行を行います。

CLAIREは、文字列の比較／類似（Jaccard、データプロファイリングから抽出した宣言的ルール、安定化したデータセット（人口ファイル、ニックネーム、セマンティック比較など））と例外を処理するためのユーザー定義ルールの両方を使用します。これらの宣言的ルールにより、ギャップや例外に対応して、能動学習のトレーニングプロセスを促進（学習に必要な照合ペアの数を削減）できるだけでなく、AIルールの特徴を迅速に作成したり、照合の精度を改善したりできます。例えば、名前、生年月日、社会保障番号の一致率が高い場合は、ルールによって一致として分類します。宣言的ルールとAIルールを組み合わせることで、トレーニングを促進して、照合の精度を改善することができます。

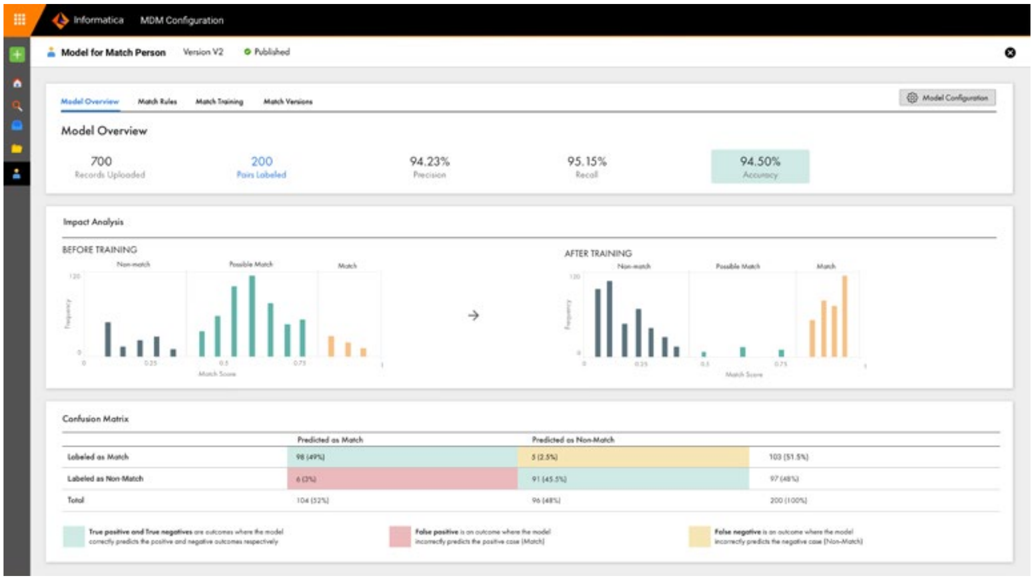

図12 : エンティティ照合

CLAIREを活用したデータガバナンスおよびコンプライアンス

今日の複雑なデータガバナンスタスク（データの検索、品質の評価、コラボレーションによるガバナンス）を自動化するためには、AIと機械学習が不可欠です。CLAIREにより、ポリシールール（データ品質など）を自動的に生成して、業務セマンティクスと技術メタデータを結び付け、ビジネスニーズに応じた最も関連性と信頼性の高いデータをユーザーに提供することができます。

データ品質のエンリッチ化を自動化

CLAIREは、Stanford NERベースのNLPアプローチを用いて、非構造化テキスト内のエンティティの解析と抽出を行います。通常、文字列（製品コードなど）からエンティティを抽出するには、ユーザーが参照テーブルや正規表現を使用して解析ルールを作成しなければなりませんが、データの量や複雑性、パターンは絶えず増えているため、すべての入力データを照合するためのルールを作成するのは非現実的であり、拡張性もありません。そこで、CLAIREではトレーニング済みのモデルを使用して、Stanford NERに基づいてエンティティを特定し、抽出します。

CLAIREは機械学習を用いて、言語、製品タイプ、技術サポート問題などの入力テキストを分類します。ここで使用する機械学習手法を、ナイーブベイズ法と最大エントロピー（多項ロジスティック回帰）法を用いた教師あり学習と呼びます。教師あり学習を使用して、モデルのトレーニングとラベル付けを行います。その後、データ処理の際にトレーニング済みのモデルを使用して、各クラスの入力テキストのラベル付け、ルーティング、処理を実行できます。例えば、類似する意味を持つテキストについて、「エンジンの問題」と「設定の問題」を区別して対応したり、複数の意味を持つテキストを使用方法に応じて区別したりするなどです。CLAIREは、製品分類と画像メタタグの抽出のためのMLモデルとNLPを活用して、画像のタグ付けと分類を自動化します。

ある世界的な大手医療機関では、フルタイムの従業員が2か月をかけて、21,000の技術資産を6,000のビジネス用語にマッピングしていました。Axon Data GovernanceとEnterprise Data Catalogを使用すれば、CLAIREにより、99%の精度で、18,000の技術資産を8分間で自動的にマッピングできます。

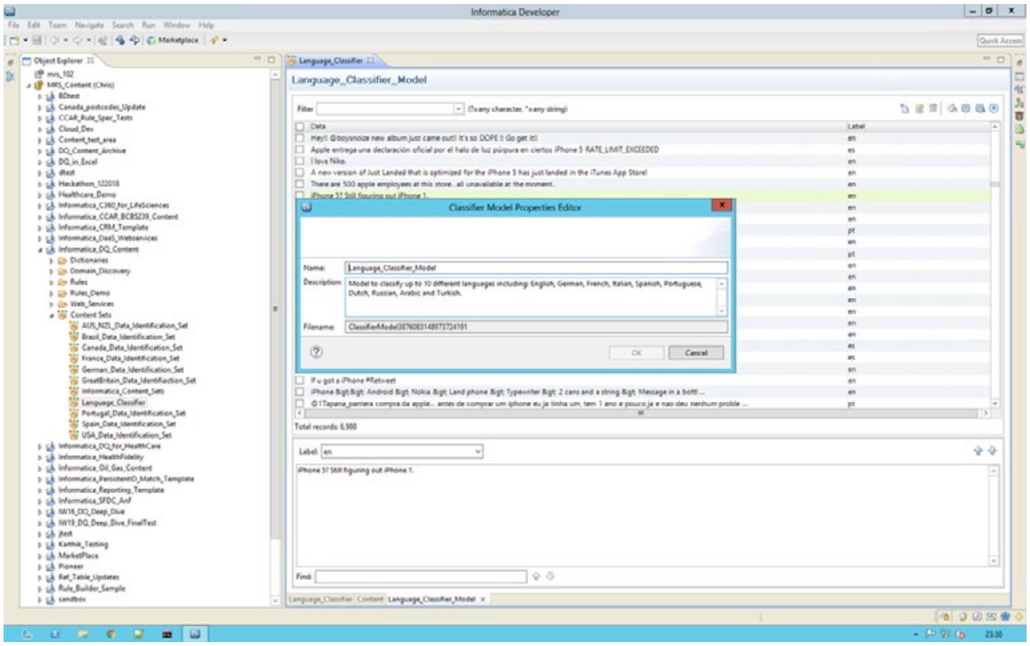

図13：機械学習のNLPがテキストを分類して、エンティティを抽出

ビジネス用語と物理データセットの関連付けを自動化

データガバナンスでは、業務上のアーチファクト、定義、利害関係者、プロセス、ポリシーなどを文書化しなければなりません。完全に整合性のあるビューを確立するためには、データ環境内の定義やビジネスビューを、該当する技術的実装と関連付ける必要があります。通常、この作業は多くの時間と労力を必要とします。また、主な担当者がコミュニケーションを図りながら、技術的な表現をひとつずつ手作業で調整しなければならないため、ミスが発生しやすくなります。完了するのに数日、数週間、または数か月かかる場合もあります。

Informatica Enterprise Data Catalogと緊密に統合されたInformatica Axon Data Governanceにより、このプロセスの近道を得ることができます。メタデータのスキャンが完了すると、リンクすべき重要で適切なデータ要素をCLAIREが提案します。これにより、データ要素の検索、検証、関連付けのタスクを簡素化できるため、データスチュワードとデータガバナンスオフィスは、より重要なタスクに専念できます。実装の進展具合によっては、このプロセスを完全に自動化することも可能です。

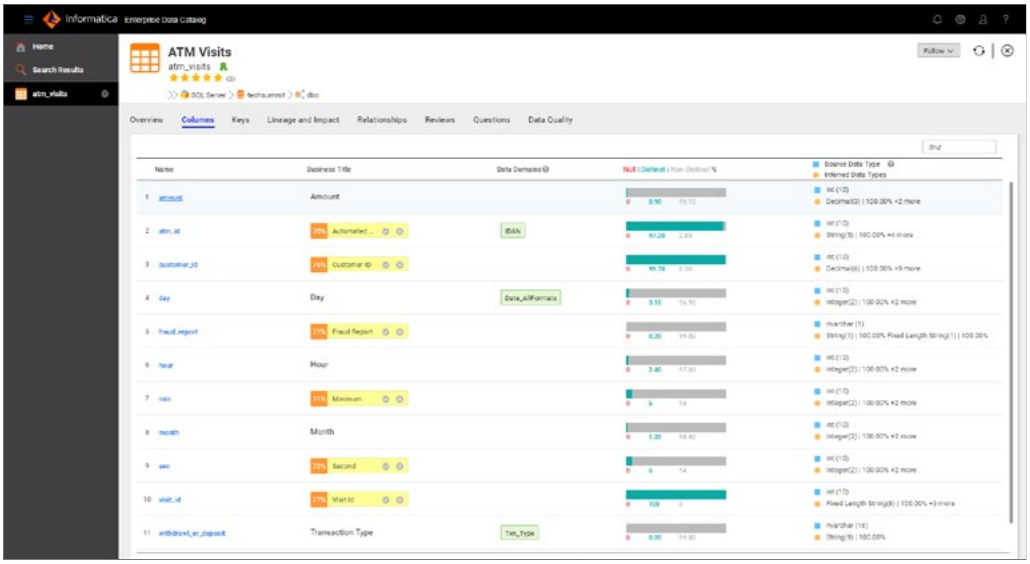

図14 : ビジネス用語と物理データセットの関連付けを自動で実行

データ品質の評価を自動化

データガバナンスでは、プロセスをサポートし、ポリシーを支えるシステム全体のデータの品質が主要業績指標（KPI）となります。データガバナンスオフィスは、データの完全性、正確性、一貫性、妥当性などを確保する必要があります。つまり、データは業務の運用をサポートするのに足る信頼性と品質を備えていなければなりません。データガバナンスの拡大に伴い、データ環境内のシステムや項目（データベースやデータレイクなど）の数が増加します。そして、それらの品質を評価するのに、ますます多くの時間が必要になります。

CLAIRE、Axon Data Governance、Informatica Data Quality、Informatica Enterprise Data Catalogを使用すれば、社内全体のデータ品質評価を自動化して、数千時間分もの作業を削減できます。データガバナンスチームが、データ品質基準ごとに、データ品質ルールをビジネス用語や重要なデータ要素に関連付けます。すると、基盤システムが各システムのデータ品質チェックを生成し、測定結果をガバナンスオフィスに報告します。

次の3つの情報を組み合わせることで、このような自動化が可能になります。

1. 重要なビジネス要素とデータ品質ルールに関する知識（Axon）
2. 移植可能／実行可能なデータ品質ルールと柔軟な実行エンジン（Informatica Data Quality）
3. 物理データ資産のメタデータ詳細（Enterprise Data Catalog）

CLAIREは、これらの情報を組み合わせることで、物理データ資産（Enterprise Data Catalog）に対するデータ品質ルールの実行ジョブを生成します（Informatica Data Quality）。また、CLAIREでは業務担当者のコンテキストも維持されるので（Axon）、適切なダッシュボードおよび集計ビューに結果を表示して、ガバナンスオフィスがその結果を利用できます。

このような自動化により、ガバナンスプログラムをより迅速に拡張できます。また、データ品質評価の作成とガバナンスコンテキストとの関連付けをひとつずつ行う必要がないため、数千時間分にも及ぶ手作業を削減できます。さらに、CLAIREは、新しい物理データ資産を識別すると、その品質を自動的に評価します。加えて、データ品質ルールの固有表現抽出／分類を用いて、新しいドメインも検出します。

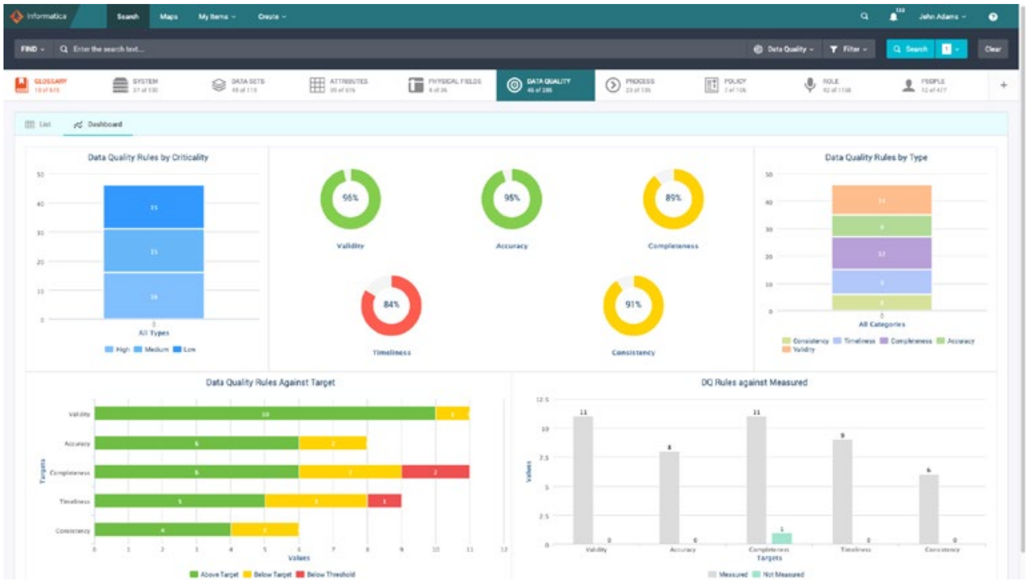

図15：データ環境全体のデータ品質評価を自動化して、数千時間分にも及ぶ手作業を削減

ML/NLPを活用したデータ品質ルールの特定

データガバナンスプログラムにはデータ品質が不可欠です。大規模なプログラムでは、多数のデータ品質ルールを使用する場合があります。CLAIREにより、データスチュワードは使用すべきルールを簡単に特定できます。不足しているルールがある場合は、CLAIREが生成します。

Axon Data Governanceを使用すれば、普通のテキストでルールの要件を指定して（例：顧客識別子は「C」で始まる8文字でなければならない）、CLAIREを呼び出すことができます。CLAIREはMLとNLPを活用して、ユーザーが指定した要件を分析し、それを技術的表現へと変換します。この表現および関連メタデータ（例：グロッサリ用語名）に基づいて、CLAIREはInformatica Data Qualityのルールを検索して、候補となるデータ品質ルールを特定します。ユーザーは、既存のルールの中から選ぶか、該当するルールがない場合は新しいデータ品質ルールを生成するようCLAIREにリクエストできます。

該当するルールがない場合は、CLAIREがInformatica Data Qualityのリポジトリ内に要件を満たす新しいデータ品質ルールを自動的に生成し、Axon Data Governanceのコンテキストにリンクします。さらに、CLAIREは、Microsoft Common Data Model (CDM) およびSalesforceのソースに基づいて、データ品質ルールをクラウドプロファイルに自動的に関連付けます。ユーザーが、上記のいずれかのソースでコアオブジェクトの新規プロファイルを作成すると、品質評価に使用すべき最適なデータ品質ルールがCLAIREによって自動的に提案されます。

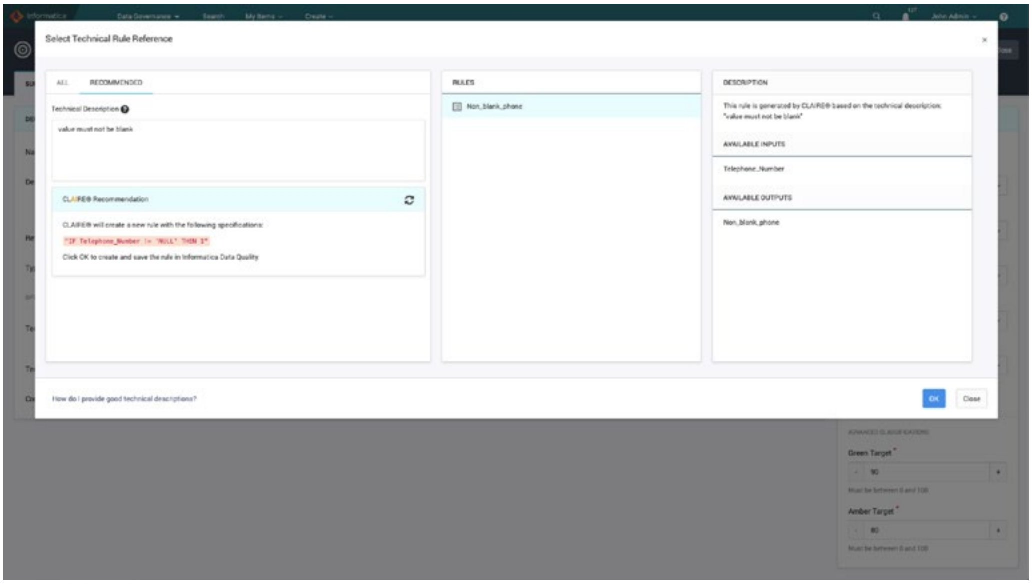

図16 : NLPを活用したデータ品質ルールの自動特定

CLAIREを活用したデータプライバシーとデータ保護

CLAIREを搭載したインテリジェントなデータプライバシーソリューションを使用すれば、データ資産に含まれている個人を特定できる情報 (PII) を全社レベルで把握して分析できます。AI主導の自動化により、個人データや機密データの探索、データ移動の把握、IDの結び付け、リスクの分析、問題の修正が実現します。

データ主体レジストリのIDマッピング

CLAIREは、IDと機密データの相互関係を判断して、データをマッピングします。これにより、プライバシーへのコンプライアンスとデータ主体アクセスレポートに対応できます。CLAIREは、組み合わせることでデータ主体を特定できるデータを評価してスコア化します。完全一致に加え、さまざまな高度な手法 (NERなど) を用いて、異なるソースのデータを組み合わせた際の結果を改善します。

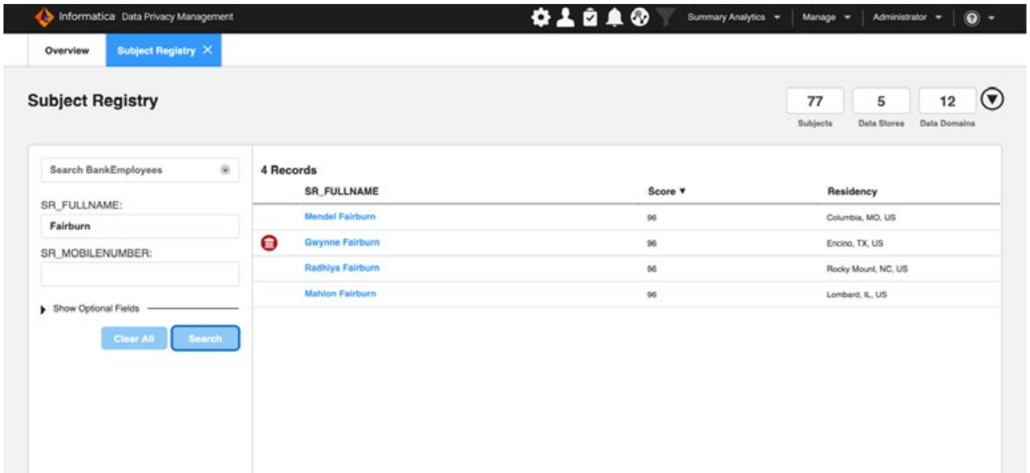

図17：データ主体レジストリのIDマッピングにより、プライバシーへのコンプライアンスやデータ主体アクセスレポートに対応

機密データのマッピングと移動

CLAIREは、上記のリネージ機能を活用および拡張して、各リポジトリでの機密データの蔓延を特定し、セキュリティ／プライバシーのコンプライアンス要件をサポートします。CLAIREは、上流と下流の両方の移動を判断し、データの特定の種類、プロセス、保護状態、データの場所など、関連するメタデータと併せて、違反が発生したかどうかを評価します。例えば、個人データが国境を超えてソースからターゲットに移動した場合や、請求プロセスのためにオンボーディングしたデータが他の部門や場所にも置かれてマーケティングプロセスに使用されているような場合、プライバシー規制に違反している可能性があります。その場合、ポリシーやプロセスの利害関係者に通知が送信され、是正措置を実行できます。

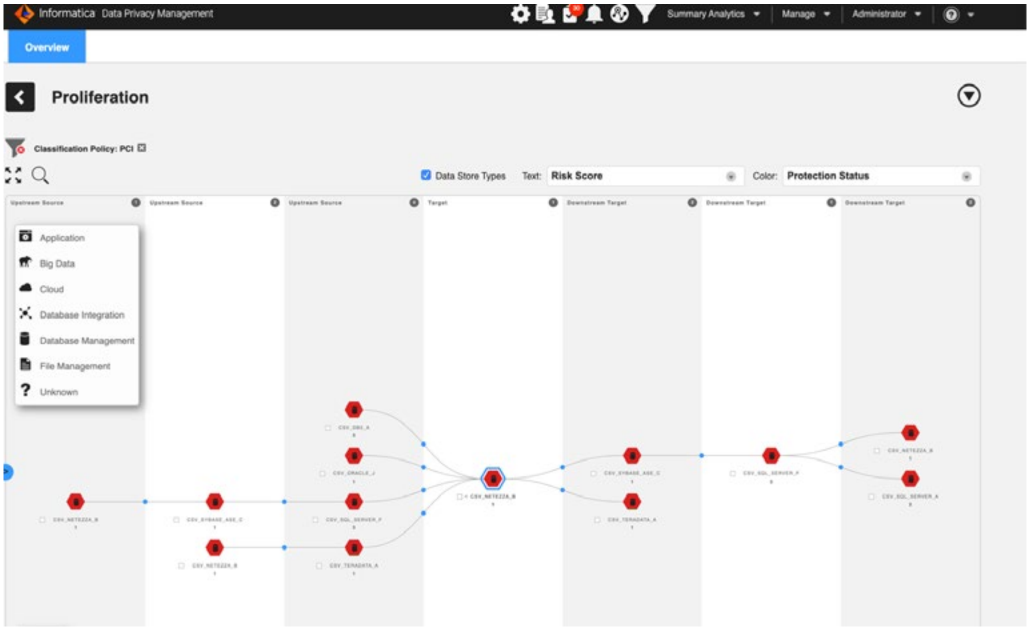

図18：機密データのリポジトリ間での移動を特定および追跡

リスクシミュレーション計画

プライバシー規制の強化により、データ保護計画の必要性が増しています。CLAIREにより、データ保護計画の効果をシミュレートして、ROI（投資収益率）を高め、予算プロセスを促進できます。CLAIREは、1つまたは複数のデータドメインに適用する保護手法を評価して、期待効用モデルに基づいて、各データストアのリスクスコア、機密データのエクスポージャ、およびその他のリスクコストの変動と全体的な効果を計算します。

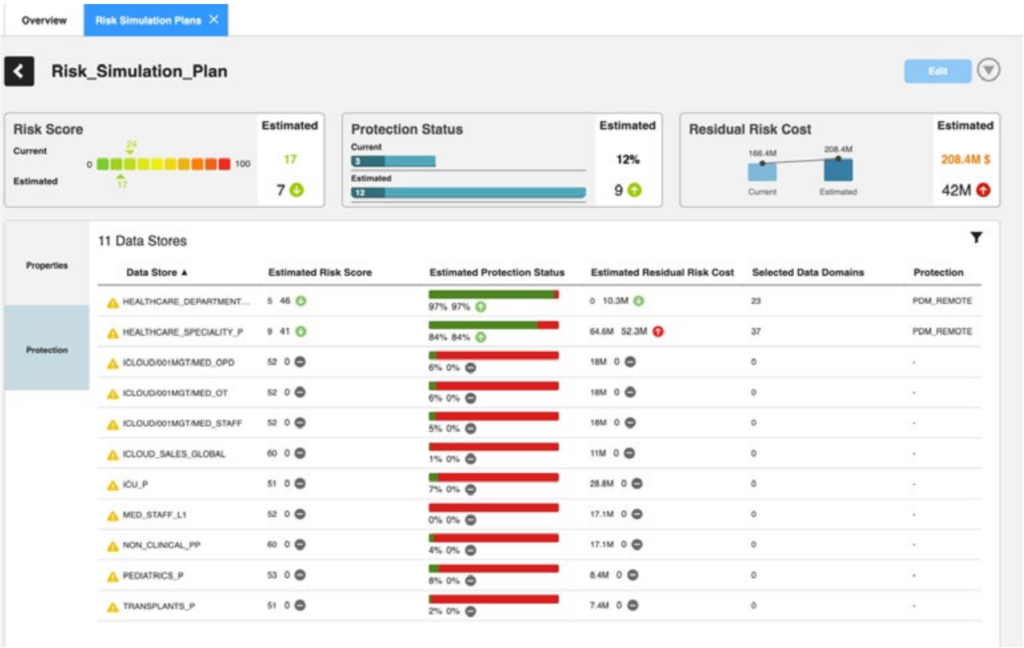

図19 : CLAIREはデータドメインに適用する保護手法を評価して、リスクを判断

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。

図20 : ユーザー行動アナリティクスにより、データの誤用を示すユーザーの異常な挙動を自動検出

リアルタイムのAPIデータ保護

APIにおける個人データの漏えいを特定して、データをブロックし、マスキングすることで、機密データ（PIIなど）をリアルタイムで保護します。Informatica API Managementにはデータ保護ライブラリが組み込まれており、APIコールを送受信する際に機密データをブロックできるため、機密データの漏えいリスクを最小限に抑えることができます。

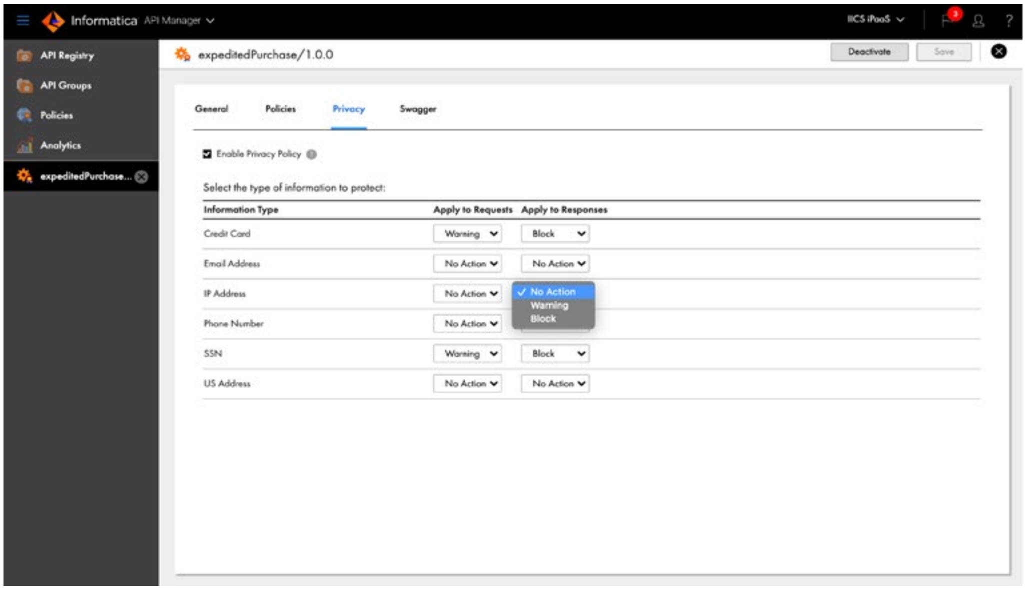

図21 : APIコールを送受信する際の機密データへのアクセスをブロック

CLAIREを活用したDataOps

CLAIREにより、さまざまなデータ管理作業を自動化して、データ処理パイプラインを高速化し、DataOpsに関連する継続的統合（CI）と継続的配信（CD）を実現できます。

豊富なインサイトに基づく、データ管理環境の予測的アナリティクス

運用アナリティクスにより、既存のプロジェクトおよびリソースの使用状況を把握して、将来のキャパシティを計画できます。チャージバックモデルを構築するためのパラメーターを獲得し、単一のデータマネジメントプラットフォームで複数の業務部門をサポートできます。リソース使用率のトレンドを継続的に観測し、これに基づいてデータ処理量を予測できるため、キャパシティプランニングに役立ちます。さらに、CLAIREは、データ管理のランタイムリソースを自動で拡張します。

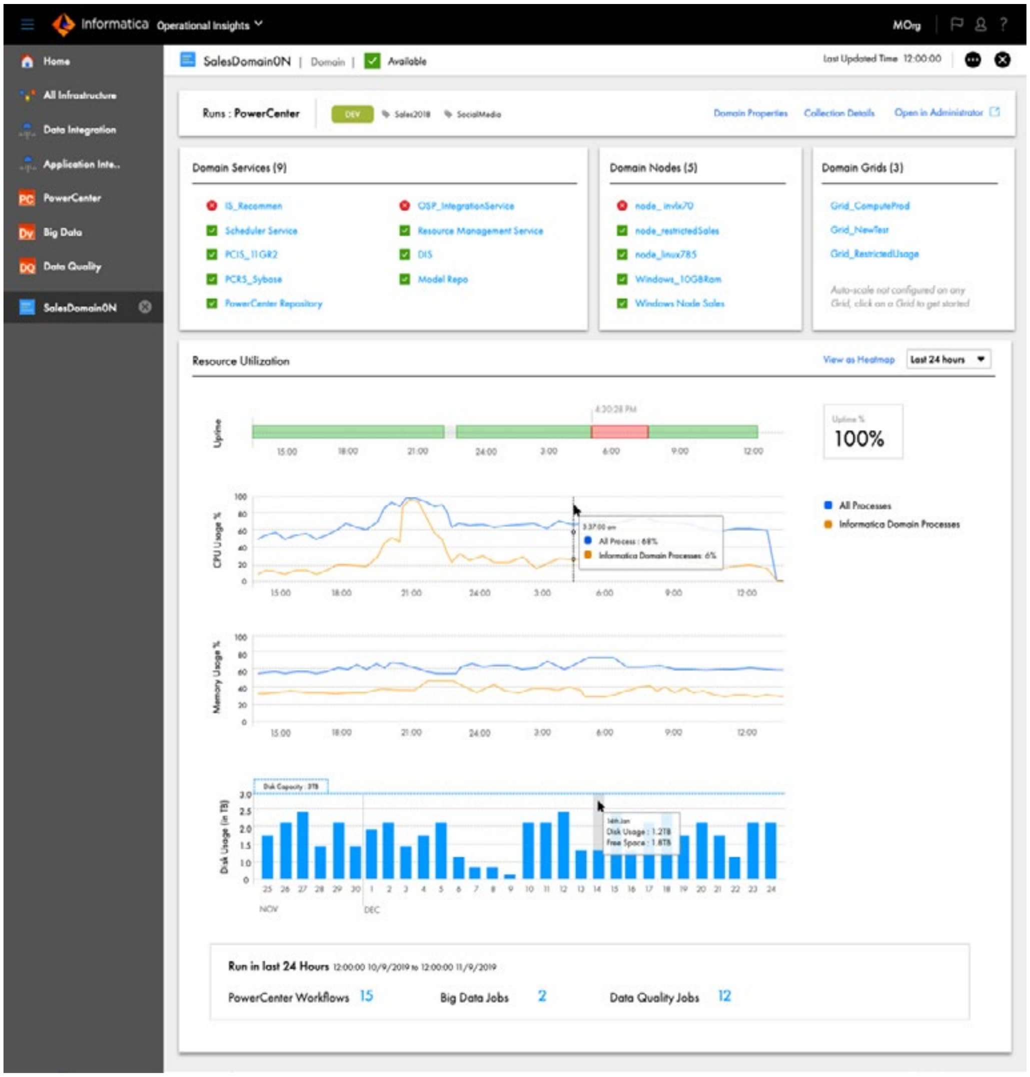

図22 : Operational Insightsでインフォマティカのドメインプロセスのリソース使用率を表示

ジョブ実行の異常検出

CLAIREは、ジョブのランタイム、データの処理、データのロード、リソースの使用、スループットなどに関する異常を自動検出します。これらの異常の自動検出により、IT部門は下流のビジネスプロセスに影響を及ぼす前にデータ統合ジョブの問題をプロアクティブに修正できます。ジョブ実行の異常検出には、Seasonal Hybrid ESDアルゴリズムを使用します。このアルゴリズムは、季節的要因（月末の負荷のピーク、ホリデーシーズンなど）を考慮に入れて、ビジネスサイクルに起因して予期される異常を含むジョブを除外します。

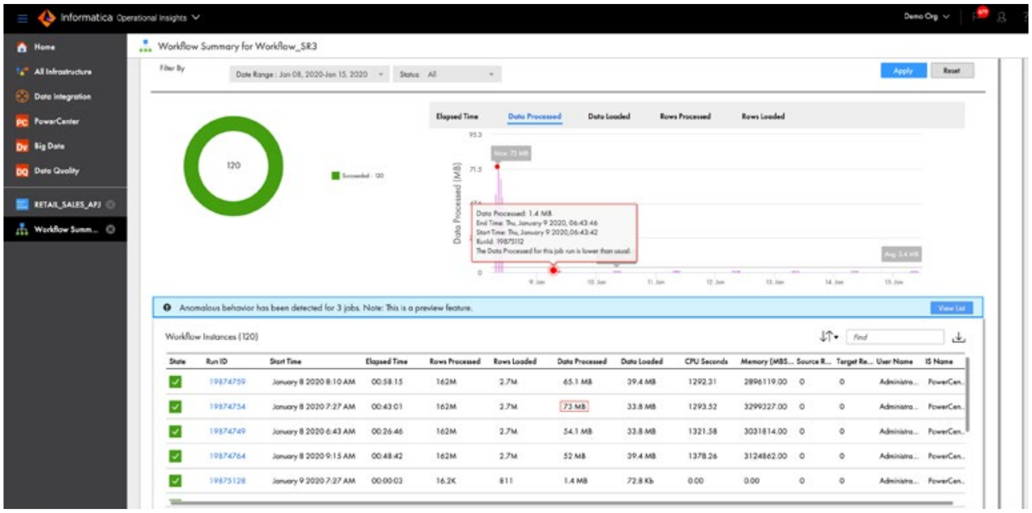

図23 : CLAIREがインフォマティカのジョブとデータ処理に関連する異常を自動検出

CLAIREの未来

CLAIREの発展に伴い、生産性と効率性が向上し、データリーダーはインテリジェントな自動化を活用して、より迅速に、より優れたインサイトを獲得できるようになります。その結果、より効果的なデータ管理を実現します。将来的に予定している機能は、次のとおりです。

1. **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。
2. **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
3. **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。
4. **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。
5. **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いて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、パフォーマンス基準を満たします。
6. **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。

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方法の1つが、エンドツーエンドのデータマネジメント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 Management Cloudによって、データの価値を最大限に引き出して活用する方法について、詳しくは当社までお問い合わせください。

Informatica

〒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

IN09_0521_03328

© Copyright Informatica LLC 2020.2021.Informatica、Informaticaロゴ、CLAIRE、Intelligent Data Management Cloud、およびAXONは、米国およびその他の国におけるInformatica LLCの商標または登録商標です。インフォマティカの商標の最新版は、https://www.informatica.com/jp/trademarks.htmlをご覧ください。その他すべての企業名および製品名は、各社が所有する商号または商標です。本文書に記載されている情報は、予告なく変更されることがあり、現状のまま提供され、明示または黙示を問わず一切の保証を伴いません。