(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第6843882号
(P6843882)
(45)発行日　令和3年3月17日(2021.3.17)　(24)登録日　令和3年2月26日(2021.2.26)

(51)Int.Cl.　F I
G06F 16/2457 (2019.01)　G06F 16/2457

請求項の数 21　(全 32 頁)

(21)出願番号　特願2018-555888(P2018-555888)
(86)(22)出願日　平成29年4月26日(2017.4.26)
(65)公表番号　特表2019-519027(P2019-519027A)
(43)公表日　令和1年7月4日(2019.7.4)
(86)国際出願番号　PCT/US2017/029583
(87)国際公開番号　WO2017/189693
(87)国際公開日　平成29年11月2日(2017.11.2)
審査請求日　令和2年4月25日(2020.4.25)
(31)優先権主張番号　15/139,186
(32)優先日　平成28年4月26日(2016.4.26)
(33)優先権主張国・地域又は機関
米国(US)

早期審査対象出願

(73)特許権者　515160389
インフォマティカ　エルエルシー
アメリカ合衆国　カリフォルニア州　９４０６３、レッドウッド　シティ、シーポート　ブルバード　２１００
２１００　Ｓｅａｐｏｒｔ　Ｂｌｖｄ，Ｒｅｄｗｏｏｄ　Ｃｉｔｙ，ＣＡ　９４０６３　Ｕ．Ｓ．Ａ．
(74)代理人　100120662
弁理士　川上　桂子
(74)代理人　100140327
弁理士　大塚　千秋

最終頁に続く

(54)【発明の名称】履歴ログからの学習と、ＥＴＬツール内のデータアセットに関するデータベースオペレーションの推奨

(57)【特許請求の範囲】
【請求項１】
１つ以上のコンピューティングデバイスによって、データ解析アプリケーションのインスタンスのユーザに推奨を提供する方法であって、
前記１つ以上のコンピューティングデバイスの少なくとも１つによって、コンテキストデータから、データベースオペレーション履歴データエントリおよびトレーニングコンテキストデータエントリをキャプチャすることによって、コンテキストデータのプロファイリングを行うステップであって、前記コンテキストデータは前記データ解析アプリケーション内のテーブル上で実行されるデータベースオペレーションに応答して前記データ解析アプリケーションの１つまたは複数のインスタンスから受信されたエントリを含む、ステップと、
前記１つ以上のコンピューティングデバイスの少なくとも１つによって、第１のセットのユーザについて、複数のテーブル上で実行される複数のデータベースオペレーションに対してプロファイリングされたデータベースオペレーション履歴データおよびプロファイリングされたコンテキストデータを維持するステップと、
前記１つ以上のコンピューティングデバイスの少なくとも１つによって、データ解析アプリケーションの第２のセットのユーザに、少なくとも１つのデータベースオペレーションまたは少なくとも１つのオペランド、の１つ以上を推薦するよう構成された複数の予測モデルを生成するステップであって、前記予測モデルのそれぞれは、プロファイルされたコンテキストデータからのコンテキストデータフィールドに対応する複数の特徴と、推薦

のための複数の対応するデータベースオペレーションまたは複数のオペランドのいずれかとを含む、ステップと、
アプリケーションコンテキストデータを含むアプリケーションログエントリを受信するステップであって、前記アプリケーションログエントリはデータ解析アプリケーションのインスタンス内のテーブル内の列を選択する第２のセットのユーザに応答して受信される、ステップと、
前記１つ以上のコンピューティングデバイスの少なくとも１つによって、前記アプリケーションコンテキストデータに少なくとも部分的に基づいて、前記複数の予測モデル内の１つ以上の予測モデルを選択するステップと、
前記１つ以上のコンピューティングデバイスの少なくとも１つによって、前記アプリケーションコンテキストデータを前記１つ以上の選択された予測モデルへ入力することにより、１つ以上の確率リストを生成するステップであって、前記確率リストのそれぞれは、前記複数のデータベースオペレーションまたは前記複数のオペランドに関連付られた複数の確率値を含む、ステップと、
前記予測モデルへの入力として使用できるフォーマットでアプリケーションコンテキストデータをキャプチャするためにアプリケーションログエントリのプロファイリングを行うステップと、
前記予測モデルへの入力として前記アプリケーションコンテキストデータを使用して、１つまたは複数の推奨データベースオペレーションを決定するステップと、
前記１つ以上のコンピューティングデバイスの少なくとも１つによって、前記１つ以上の確率リストに少なくとも部分的に基づいて、１つ以上の推奨を決定するステップであって、前記１つ以上の推奨における各推奨は、データベースオペレーションまたはオペランドを含む、ステップと、
前記１つ以上のコンピューティングデバイスの少なくとも１つによって、ユーザへの提示のためにデータ解析アプリケーションのインスタンスに前記１つ以上の推奨を送信するステップと、を含む方法。
【請求項２】
前記複数の予測モデルを生成するステップが、複数の予測モデルのそれぞれについて、プロファイルされたコンテキストデータから複数のコンテキストデータフィールドを選択することにより複数の特徴を決定することと、
推薦する複数のデータベースオペレーションまたは複数のオペランドを決定することと、
複数のデータベースオペレーションまたは複数のオペランドの各々について、複数の特徴の各々についての特徴重みを決定することとを含み、
前記特徴重みはデータベースオペレーションまたはオペランドに関する特徴の予測性の尺度に対応する、請求項１に記載の方法。
【請求項３】
前記コンテキストデータは、プロジェクトメタデータ、ワークシートメタデータ、およびユーザメタデータのうちの少なくとも１つを含む、請求項１に記載の方法。
【請求項４】
少なくとも１つの予測モデルが、多項ロジスティック分類器である、請求項１に記載の方法。
【請求項５】
前記アプリケーションコンテキストデータは、プロジェクトメタデータ、ワークシートメタデータ、およびユーザメタデータのうちの少なくとも１つを含む、請求項１に記載の方法。
【請求項６】
前記１つ以上の推奨は、ジョインオペレーションおよび結合オペレーションのうちの少なくとも１つを含む、請求項１に記載の方法。
【請求項７】

前記１つ以上の選択された予測モデルは、オペレーションモデルおよびオペランドモデルを含み、
前記アプリケーションコンテキストデータを前記１つ以上の選択された予測モデルへ入力することにより、１つ以上の確率リストを生成するステップは、
前記アプリケーションコンテキストデータを前記オペレーションモデルへ入力することにより、前記複数のデータベースオペレーションに関連付られた確率を含む第１の確率リストを生成することと、
前記アプリケーションコンテキストデータおよび前記第１の確率リストを前記オペランドモデルへ入力することにより、前記複数のオペランドに関連付られた確率を含む第２の確率リストを生成することと、を含む、
請求項１に記載の方法。
【請求項８】
データ解析アプリケーションのインスタンスのユーザに推奨を提供する装置であって、
１つ以上のプロセッサと、
前記１つ以上のプロセッサの少なくとも１つに動作可能に結合された１つ以上のメモリとを備え、
前記１つ以上のメモリは、前記１つ以上のプロセッサの少なくとも１つに実行されたときに、前記１つ以上のプロセッサの少なくとも１つに、
コンテキストデータから、データベースオペレーション履歴データエントリおよびトレーニングコンテキストデータエントリをキャプチャすることによって、コンテキストデータのプロファイリングを行うステップであって、前記コンテキストデータは前記データ解析アプリケーション内のテーブル上で実行されるデータベースオペレーションに応答して前記データ解析アプリケーションの１つまたは複数のインスタンスから受信されたエントリを含む、ステップと、
第１のセットのユーザについて、複数のテーブル上で実行される複数のデータベースオペレーションに対してプロファイリングされたデータベースオペレーション履歴データおよびプロファイリングされたコンテキストデータを維持するステップと、
データ解析アプリケーションの第２のセットのユーザに、少なくとも１つのデータベースオペレーションまたは少なくとも１つのオペランド、の１つ以上を推薦するよう構成された複数の予測モデルを生成するステップであって、前記予測モデルのそれぞれは、プロファイルされたコンテキストデータからのコンテキストデータフィールドに対応する複数の特徴と、推薦のための複数の対応するデータベースオペレーションまたは複数のオペランドのいずれかとを含む、ステップと、
アプリケーションコンテキストデータを含むアプリケーションログエントリを受信するステップであって、前記アプリケーションログエントリはデータ解析アプリケーションのインスタンス内のテーブル内の列を選択する第２のセットのユーザに応答して受信される、ステップと、
前記アプリケーションコンテキストデータに少なくとも部分的に基づいて、前記複数の予測モデル内の１つ以上の予測モデルを選択するステップと、
前記アプリケーションコンテキストデータを前記１つ以上の選択された予測モデルへ入力することにより、１つ以上の確率リストを生成するステップであって、前記確率リストのそれぞれは、前記複数のデータベースオペレーションまたは前記複数のオペランドに関連付られた複数の確率値を含む、ステップと、
前記１つ以上の確率リストに少なくとも部分的に基づいて、１つ以上の推奨を決定するステップであって、前記１つ以上の推奨における各推奨は、データベースオペレーションまたはオペランドを含む、ステップと、
ユーザへの提示のためにデータ解析アプリケーションのインスタンスに前記１つ以上の推奨を送信するステップと、
を実行させる命令が格納されている、装置。
【請求項９】

前記命令は、前記１つ以上のプロセッサの少なくとも１つに実行されたときに、前記１つ以上のプロセッサの少なくとも１つに、
前記複数の予測モデルを生成するステップにおいて、複数の予測モデルのそれぞれについて、
プロファイルされたコンテキストデータから複数のコンテキストデータフィールドを選択することにより複数の特徴を決定することと、
推薦する複数のデータベースオペレーションまたは複数のオペランドを決定することと、
複数のデータベースオペレーションまたは複数のオペランドの各々について、複数の特徴の各々についての特徴重みを決定することと、をさらに行わせ、
前記特徴重みはデータベースオペレーションまたはオペランドに関する特徴の予測性の尺度に対応する、請求項８に記載の装置。
【請求項１０】
前記コンテキストデータは、プロジェクトメタデータ、ワークシートメタデータ、およびユーザメタデータのうちの少なくとも１つを含む、請求項８に記載の装置。
【請求項１１】
少なくとも１つの予測モデルが、多項ロジスティック分類器である、請求項８に記載の装置。
【請求項１２】
前記アプリケーションコンテキストデータは、プロジェクトメタデータ、ワークシートメタデータ、およびユーザメタデータのうちの少なくとも１つを含む、請求項８に記載の装置。
【請求項１３】
前記１つ以上の推奨は、ジョインオペレーションおよび結合オペレーションのうちの少なくとも１つを含む、請求項８に記載の装置。
【請求項１４】
前記１つ以上の選択された予測モデルは、オペレーションモデルおよびオペランドモデルを含み、
前記命令は、前記１つ以上のプロセッサの少なくとも１つに実行されたときに、前記１つ以上のプロセッサの少なくとも１つに、
前記アプリケーションコンテキストデータを前記１つ以上の選択された予測モデルへ入力することにより、１つ以上の確率リストを生成するステップにおいて、
前記アプリケーションコンテキストデータを前記オペレーションモデルへ入力することにより、前記複数のデータベースオペレーションに関連付られた確率を含む第１の確率リストを生成することと、
前記アプリケーションコンテキストデータおよび前記第１の確率リストを前記オペランドモデルへ入力することにより、前記複数のオペランドに関連付られた確率を含む第２の確率リストを生成することと、をさらに実行させる、
請求項８に記載の装置。
【請求項１５】
コンピュータ可読命令を記憶する少なくとも１つの非一時的なコンピュータ可読記憶媒体であって、前記コンピュータ可読命令は、１つ以上のコンピューティングデバイスに実行されたとき、前記１つ以上のコンピューティングデバイスの少なくとも１つに、
コンテキストデータから、データベースオペレーション履歴データエントリおよびトレーニングコンテキストデータエントリをキャプチャすることによって、コンテキストデータのプロファイリングを行うステップであって、前記コンテキストデータはデータ解析アプリケーション内のテーブル上で実行されるデータベースオペレーションに応答して前記データ解析アプリケーションの１つまたは複数のインスタンスから受信されたエントリを含む、ステップと、
第１のセットのユーザについて、複数のテーブル上で実行される複数のデータベースオ

ペレーションに対してプロファイリングされたデータベースオペレーション履歴データおよびプロファイリングされたコンテキストデータを維持するステップと、
データ解析アプリケーションの第２のセットのユーザに、少なくとも１つのデータベースオペレーションまたは少なくとも１つのオペランド、の１つ以上を推薦するよう構成された複数の予測モデルを生成するステップであって、前記予測モデルのそれぞれは、プロファイルされたコンテキストデータからのコンテキストデータフィールドに対応する複数の特徴と、推薦のための複数の対応するデータベースオペレーションまたは複数のオペランドのいずれかとを含む、ステップと、
アプリケーションコンテキストデータを含むアプリケーションログエントリを受信するステップであって、前記アプリケーションログエントリはデータ解析アプリケーションのインスタンス内のテーブル内の列を選択する第２のセットのユーザに応答して受信される、ステップと、
前記アプリケーションコンテキストデータに少なくとも部分的に基づいて、前記複数の予測モデル内の１つ以上の予測モデルを選択するステップと、
前記アプリケーションコンテキストデータを前記１つ以上の選択された予測モデルへ入力することにより、１つ以上の確率リストを生成するステップであって、前記確率リストのそれぞれは、前記複数のデータベースオペレーションまたは前記複数のオペランドに関連付られた複数の確率値を含む、ステップと、
前記１つ以上の確率リストに少なくとも部分的に基づいて、１つ以上の推奨を決定するステップであって、前記１つ以上の推奨における各推奨は、データベースオペレーションまたはオペランドを含む、ステップと、
ユーザへの提示のためにデータ解析アプリケーションのインスタンスに前記１つ以上の推奨を送信するステップと、
を実行させる、記憶媒体。
【請求項１６】
前記コンピュータ可読命令は、前記１つ以上のプロセッサの少なくとも１つに実行されたときに、前記１つ以上のプロセッサの少なくとも１つに、
前記複数の予測モデルを生成するステップにおいて、複数の予測モデルのそれぞれについて、
プロファイルされたコンテキストデータから複数のコンテキストデータフィールドを選択することにより複数の特徴を決定することと、
推薦する複数のデータベースオペレーションまたは複数のオペランドを決定することと、
複数のデータベースオペレーションまたは複数のオペランドの各々について、複数の特徴の各々についての特徴重みを決定することと、をさらに行わせ、
前記特徴重みはデータベースオペレーションまたはオペランドに関する特徴の予測性の尺度に対応する、請求項１５に記載の記憶媒体。
【請求項１７】
前記コンテキストデータは、プロジェクトメタデータ、ワークシートメタデータ、およびユーザメタデータのうちの少なくとも１つを含む、請求項１５に記載の記憶媒体。
【請求項１８】
少なくとも１つの予測モデルが、多項ロジスティック分類器である、請求項１５に記載の記憶媒体。
【請求項１９】
前記アプリケーションコンテキストデータは、プロジェクトメタデータ、ワークシートメタデータ、およびユーザメタデータのうちの少なくとも１つを含む、請求項１５に記載の記憶媒体。
【請求項２０】
前記１つ以上の推奨は、ジョインオペレーションおよび結合オペレーションのうちの少なくとも１つを含む、請求項１５に記載の記憶媒体。

【請求項２１】
前記１つ以上の選択された予測モデルは、オペレーションモデルおよびオペランドモデルを含み、
前記コンピュータ可読命令は、前記１つ以上のプロセッサの少なくとも１つに実行されたときに、前記１つ以上のプロセッサの少なくとも１つに、
前記アプリケーションコンテキストデータを前記１つ以上の選択された予測モデルへ入力することにより、１つ以上の確率リストを生成するステップにおいて、
前記アプリケーションコンテキストデータを前記オペレーションモデルへ入力することにより、前記複数のデータベースオペレーションに関連付られた確率を含む第１の確率リストを生成することと、
前記アプリケーションコンテキストデータおよび前記第１の確率リストを前記オペランドモデルへ入力することにより、前記複数のオペランドに関連付られた確率を含む第２の確率リストを生成することと、をさらに実行させる、
請求項１５に記載の記憶媒体。
【発明の詳細な説明】
【技術分野】
【０００１】
本出願は２０１６年４月２６日に出願された米国非仮出願第１５／１３９，１８６号の優先権を主張し、その開示は、その全体が参照により本明細書に組み込まれる。
【０００２】
本開示は一般に、データベース管理システムおよびデータウェアハウスにおけるデータプロセスの抽出、変換、およびロードに関し、より詳細には、データ閲覧および編集環境において表示されるデータのためのデータベースオペレーションを決定し、推奨するためのコンピュータ実行方法に関する。
【背景技術】
【０００３】
データウェアハウスの分野では、複数の外部データソースからのデータが通常、内部データベース管理システムに取り込まれるときに、抽出（ｅｘｔｒａｃｔ）、変換（ｔｒａｎｓｆｏｒｍ）、およびロード（ｌｏａｄ）（ＥＴＬ）プロセスを介して遷移する。ＥＴＬプロセスの一部として、データは、（ｉ）１つまたは複数のデータソースから抽出され、（ｉｉ）内部データソースのビジネス要件および技術要件に従ってプログラム変換され、（ｉｉｉ）内部データベース管理システムのターゲットデータストアにロードされる。一旦システムに入ると、データは、様々なデータベースオペレーションを使用してシステムユーザによって操作され得る。多くの場合、ユーザは膨大な量のデータを扱っており、一部のユーザは、データベース管理アプリケーションがデータを処理するためにサポートするデータベースオペレーションに慣れていないか、またはデータベース管理システム内でデータを処理する最も効率的な方法を知らない。この問題に対処するのに十分な知識および経験を獲得することは、特に、一時ユーザまたは多くのタイプのデータを扱うユーザにとって困難であり、時間がかかる可能性がある。
【発明の概要】
【０００４】
データ解析サーバは、機械学習予測モデルを使用して、プログラム的に決定された推奨データベース動作を、データ解析アプリケーションの習熟度が低いユーザ（ガイド付きユーザ）に提供するように構成される。予測モデルは、データベース内の類似データに関する上級ユーザ（トレーニングユーザ）によるデータベースオペレーション入力から学習される。予測モデルは、熟練度の低いユーザがどのデータベースオペレーションがデータに適しているかを選択するプロセスを改善することによって、データベースを操作する際の効率を改善することを可能にする。
【０００５】
データ解析サーバは、以前のデータベースユーザによるデータベースオペレーションの

履歴データを使用して、ＥＴＬツールのユーザにデータベースオペレーションを推奨するための予測モデルを構築する。データプロファイリングモジュールは、選択されたユーザグループに提示され、ユーザによって操作されるデータベーステーブルおよびテーブルセット（プロジェクト）のコンテキストデータを維持するように構成される。コンテキストデータは、テーブルおよびプロジェクトのメタデータを含む。データベースオペレーション履歴モジュールは、テーブルおよびプロジェクト上のデータベースオペレーションの履歴データを維持するように構成される。本明細書で使用されるデータベースオペレーションは、ＥＴＬによってサポートされ、変換または変更されたデータセットを生成するために特定のデータに対して実行されるプログラム操作である。特定のデータベースオペレーションには、ジョイン（コンバイン）、結合（マージ）、フィルタ、フォーミュラ、ルックアップ、列分割、列追加（データ拡張）、パターン認識および不整合修正、データクレンジング、データ整合、データ標準化などが含まれる。データベース演算は、数学的演算、方程式などのデータに対する演算をさらに含むことができる。

【０００６】

データベースオペレーション推奨モジュールはデータベースオペレーションをユーザに推奨するための予測モデルを構築し、トレーニングし、使用するように構成される。データベースオペレーション推奨モジュールは維持されたデータベースオペレーション履歴データおよびコンテキストデータを使用してモデルをトレーニングし、それによって、どのコンテキストデータが特定のデータベースオペレーションの適用を予測するかを決定する。ガイド付きユーザによるデータベースの使用中に、リアルタイムでガイド付きユーザに対する推薦を生成するために、データベースオペレーション推薦モジュールはガイド付きユーザによってアクセスされている特定のテーブルまたはプロジェクトに対するコンテキストデータを受信し、予測モデルを使用してそのテーブルまたはプロジェクトに対して実行する１つまたは複数の推薦データベースオペレーションを決定する。

【０００７】

データ解析アプリケーションのグラフィカルユーザインタフェースは、データセクション、情報セクション、および様々なユーザインタフェース制御を含む。データセクションは、分析用のテーブルを表示するためのものである。情報セクションは、テーブルのスキーマ定義に基づいて、テーブルのプロファイル情報を表示するものである。コンポジットデータ制御はテーブル間の少なくとも１つのマッチング列に基づいてテーブルをコンポジットテーブルに統合するデータベースオペレーション（同等には、データベースコマンド）を受け取るためのものである。複合データ制御は、様々な統一データベースオペレーションのための複数の異なる制御であってもよい。ＵＩの推薦制御は、データベースオペレーション推薦モジュールによって決定された推薦データベースオペレーションを表示するためのものである。

【０００８】

本明細書に記載される特徴および利点はすべてを包含するものではなく、特に、多くの追加の特徴および利点が、図面、明細書、および特許請求の範囲を考慮して、当業者には明らかであろう。さらに、本明細書で使用される言語は主に、読みやすさおよび説明の目的のために選択されており、本発明の主題を描写または限定するために選択されていない場合があることに留意されたい。

【図面の簡単な説明】

【０００９】

【図１】図１は、一実施形態による、データ解析アプリケーションにおいて、データベースオペレーションの履歴ログから予測モデルを生成し、データに対するデータベースオペレーションを推奨するコンピューティング環境の高レベルブロック図である。

【図２】図２は、一実施形態によるデータベースオペレーション推薦モジュールのより詳細な図を示す。

【図３】図３は、予測モデルをトレーニングする際に使用するための特徴およびクラスを示す例示的なデータテーブルである。

【図４】図４は、一実施形態による、データ解析アプリケーションにおいてデータを閲覧および操作するためのユーザインタフェースの一例を示す。
【図５Ａ】図５Ａは一実施形態による、データ解析アプリケーションのガイド付きユーザに対してデータベースオペレーションを決定し推奨するための予測モデルを構築し、トレーニングするための方法を示すフローチャートである。
【図５Ｂ】図５Ｂは、一実施形態による、データ解析アプリケーションのガイド付きユーザにデータベースオペレーションを推薦するためにトレーニングされた予測モデルを使用する方法を示すフローチャートである。
【図６】図６は、一実施形態による、選択された列に応答して提供される推奨を備えた、図３の例示的なユーザインタフェースを示す。
【図７】図７は、データ解析アプリケーションにおいて、データ解析サーバから受信した推奨データベースオペレーションおよびオペランドを提示するための方法を示すフローチャートである。
【発明を実施するための形態】
【００１０】
システムのアーキテクチャ
図１は、一実施形態による、データ解析アプリケーションにおいて、データベースオペレーションの履歴ログから予測モデルを生成し、データに対するデータベースオペレーションを推奨するコンピューティング環境１００の高レベルブロック図である。
【００１１】
示されるように、コンピューティング環境１００は、データリポジトリ１０２、データ解析サーバ１０４、およびデータ解析アプリケーション１２５を含む。
【００１２】
複数のデータリポジトリ１０２（本明細書では個別にデータリポジトリ１０２とも呼ぶ）は、データを管理するための１つまたは複数のシステムを含む。各データリポジトリ１０２は、データリポジトリ１０２内に格納されたデータにアクセスして更新するためのチャネルを提供する。データリポジトリ１０２内のデータは、ユーザ、ユーザのグループ、エンティティ、および／またはワークフローに関連付けられ得る。例えば、データリポジトリ１０２は、特定のエンティティに関連付けられたすべての個人に関連付けられたデータを記憶する顧客関係管理（ＣＲＭ）システムまたは人事（ＨＲ）管理システムとすることができる。データリポジトリ１０２は、ＥＴＬプロセスのためのデータソースまたはエクスポートターゲットとすることができる。データソースの例は、データベース、アプリケーション、およびローカルファイルを含む。同様に、これらのソースは、データをエクスポートするためのターゲットとして機能することができる。共通のエクスポートターゲットは、ＴＡＢＬＥＡＵ、ＳＡＬＥＳＦＯＲＣＥ　ＷＡＶＥ、およびＥＸＣＥＬである。
【００１３】
データ解析アプリケーション１２５は、ユーザがデータ解析サーバ１０４によってデータリポジトリ１０２から抽出されたデータを操作し、単一のテーブル又は多数のテーブルに対して実行されるべきデータベースオペレーションを選択及び指定することを可能にするソフトウエアアプリケーションであり、この機能を実行するための１つの手段である。一実施形態では、データ解析アプリケーション１２５がテーブルのセットであるプロジェクトの形でユーザにデータを提供する。データ解析アプリケーション１２５の様々なモジュールは、汎用コンピュータシステムのネイティブコンポーネントまたは標準コンポーネントではなく、コンピュータシステムの汎用機能を超えて拡張する、本明細書で説明する特定の機能を提供する。さらに、モジュールの機能および動作はコンピュータシステムによる実装を必要とするほど十分に複雑であり、したがって、いかなる実際的な実施形態でも、人間の心の中の精神的なステップによって実行することはできない。これらの構成要素の各々は、以下により詳細に記載される。データ解析アプリケーション１２５はデバイス非依存であり、したがって、デスクトップアプリケーション、モバイルアプリケーション、またはウェブベースのアプリケーションとすることができる。その様々な機能を実行

するために、データ解析アプリケーション１２５は、ユーザインタフェース（ＵＩ）モジュール１２２およびデータベースオペレーションＵＩモジュール１２４を含む。

【００１４】

いくつかの実施形態では、データ解析アプリケーション１２５は、様々なオンサイトおよび外部のソースおよびターゲット、ならびに本明細書で説明されるプロセスに関与する強化サービスと共に、より大きなクラウドアーキテクチャの一部である。

【００１５】

ＵＩモジュール１２２はＵＩにおいて表示するためのデータを受信し、受信したデータに対応するユーザインタフェースを生成し、受信したデータをテーブルにポピュレートし、予測モデルに基づいてデータリファインメントの推奨を表示し、テーブルの１つまたは複数の列に関連付けられた列サマリを生成し、これらの機能を実行するための１つの手段である。生成されたユーザインタフェースは、データ解析アプリケーション１２５のユーザがテーブルエントリを操作すること、およびデータベースオペレーションをデータに適用することを含めて、テーブルを見ること、およびテーブルと対話することを可能にする。

【００１６】

データベースオペレーションＵＩモジュール１２４はＵＩモジュール１２２によって生成されたテーブル内のデータに適用するための１つ以上のデータベースオペレーション制御を提供し、この機能を実行するための１つの手段である。具体的には、データベースオペレーションＵＩモジュール１２４がデータ解析アプリケーション１２５のユーザがテーブルに関連付けられたデータベースオペレーションを選択し、指定し、および／またはデータベースオペレーションの適用を引き起こすことを可能にする制御を提供する。

【００１７】

一実施形態によれば、ＵＩモジュール１２２およびデータベースオペレーションＵＩモジュール１２４によって提供されるユーザインタフェースは、グラフィカルに表現されたデータセクション、情報セクション、および様々なグラフィカルに表現されたデータベースオペレーション制御を含む。ＵＩのデータ部は、解析用のテーブルを表示するためのものである。ＵＩの情報セクションは、テーブルに関するプロファイル情報を表示するためのものである。プロファイル情報は、コンテキストデータなどのテーブルの特徴を記述する。ＵＩの複合データ制御は、テーブル間の少なくとも１つの一致する列に基づいて２つのテーブルを複合テーブルに統合するコマンドを受信するユーザインタフェース要素である。ＵＩの推薦制御は、予測モデルを用いてデータベースオペレーション推薦モジュール１１４により決定された推薦データベースオペレーションを表示するユーザインタフェース要素である。ＵＩは、図４および図６に関して以下により詳細に説明される。

【００１８】

データベースオペレーションＵＩモジュール１２４は、実行された各データベースオペレーションに対して、表示されたテーブルに対して実行された各データベースオペレーションをデータ解析サーバ１０４内のデータベースオペレーション履歴モジュール１１２に送信する。各データベースオペレーションは、オペレーション識別子によって表され、オペレーション識別子は例えば、名前、ＩＤ番号、およびデータベースオペレーションに含まれていたオペランドを示すオペレーション記述によって、オペレーションを一意に識別する。データベースオペレーション履歴モジュール１１２は、データに適用されたデータベースオペレーションをデータベースオペレーション履歴記憶部１２０に記憶する。経時的にデータに適用されるデータベースオペレーションはデータベースオペレーション履歴ストア１２０に取り込まれ、データベースオペレーション履歴内の任意のステップはアンドゥ、リドゥ、または異なるデータに適用することができる。データベースオペレーションは後述するように、ログの形式で格納することができる。

【００１９】

データ解析サーバ１０４はデータをデータリポジトリ１０２から抽出し、データを処理し、処理されたデータをデータ解析アプリケーション１２５に提供して、データをユーザ

に表示し、ユーザによって操作できるようにする。これらの機能を実行するために、データ解析サーバ１０４は、データ抽出モジュール１０８と、データプロファイリングモジュール１１０と、データベースオペレーション履歴モジュール１１２とを含む。さらに、これらの機能に関連するデータを記憶するために、データプロファイリングサーバ１０４は、リポジトリデータストア１１６、プロファイリングデータストア１１８、およびデータベースオペレーション履歴ストア１２０を含む。分析サーバ１０４の様々なモジュールは、汎用コンピュータシステムのネイティブコンポーネントまたは標準コンポーネントではなく、コンピュータシステムの汎用機能を超えて拡張する、本明細書で説明する特定の機能を提供する。さらに、モジュールの機能および動作はコンピュータシステムによる実装を必要とするほど十分に複雑であり、したがって、いかなる実際的な実施形態でも、人間の心の中の精神的なステップによって実行することはできない。これらの構成要素の各々は、以下により詳細に記載される。

【００２０】

データ抽出モジュール１０８は抽出されるべきデータリポジトリ１０２内のデータを識別し、そのデータをデータリポジトリ１０２から取り出し、そのデータをリポジトリデータストア１１６に格納するように構成され、そのための１つの手段である。動作中、データ抽出モジュール１０８は、データを抽出する１つまたは複数のデータリポジトリ１０２を識別する。データ抽出モジュール１０８はまた、抽出されるべき識別されたデータリポジトリ１０２に記憶された特定のデータを識別する。データリポジトリ１０２および／またはそこに格納された特定のデータの識別は、データプロファイリング動作を行うユーザから受け取った命令に基づいて行うことができる。あるいは、そのような識別がデータを抽出する外部データソースを指定する１つまたは複数のビジネスロジック定義に基づいて行うことができる。

【００２１】

データ抽出モジュール１０８は、データリポジトリ１０２によって提供されるデータアクセスチャネルを介してデータリポジトリ１０２から識別されたデータを抽出する。一実施形態では、データ・アクセス・チャネルは、データ抽出モジュール１０８がデータ・リポジトリ１０２と安全に通信して、データ・リポジトリ１０２との間でデータを取り出し、送信することを可能にする安全なデータ転送プロトコルである。データがデータリポジトリ１０２から抽出されると、データ抽出モジュール１０８は、データをリポジトリデータストア１１６に格納する。

【００２２】

データプロファイリングモジュール１１０はデータリポジトリ１０２から抽出され、リポジトリデータストア１１６に格納されたデータを処理して、データのすべての列、行、および領域を完全にプロファイリングし、そうするための１つの手段である。列、行、およびデータフィールドのプロファイリングは、データタイプ、データドメイン、およびエントリ長、固有値パーセント、および空白値パーセントなどのデータ値に関する他の情報を識別することを含む。

【００２３】

データベース運用履歴部１１２は、セル、テーブル、プロジェクトに適用されるデータベース運用の履歴を受け取り、格納する手段の一つである。動作中、データベース動作がセル、テーブル、またはプロジェクトに適用されるとき、データベース動作履歴モジュール１１２は、適用された特定のデータベース動作と、どのデータに適用されたかを、データベース動作履歴ストア１２０に記憶する。したがって、経時的にデータに適用されるデータベースオペレーションは、データベースオペレーション履歴ストア１２０に取り込まれる。

【００２４】

本明細書で使用されるデータベースオペレーションは、ＥＴＬシステムのプログラムコードによってサポートされ、変換または変更されたデータセットを生成するために特定のデータに対して実行されるプログラム操作である。データベースオペレーションは、テー

ブルまたはプロジェクトに対して実行することができる。特定のデータベースオペレーションには、ジョイン（コンバイン）、結合（マージ）、フィルタ、フォーミュラ、ルックアップ、列分割、列追加（データ拡張）、パターン認識および不整合修正、データクレンジング、データ整合、データ標準化が含まれる。データベース演算は、数学的演算、数式などのデータに対する演算をさらに含むことができる。データベースオペレーションには以下のものがある。

【００２５】

【表１】

表１：データベースオペレーション

データベース オペレーション	説明
ジョイン（join）	２つのソース間の列の１つまたは複数のペアと一致する条件に基づいてソースを結合
結合（union）	異なるソースのデータを同じフィールド名でマージ
expr	オペランドとして与えられる数式を評価
フィルタ（filter）	テーブルまたは列の行を指定された条件に基づいてフィルタリング
ルックアップ（lookup）	別のデータベーステーブルまたはビューの値を返す
列追加（column add）	新しい列を追加（行ごとに）。新しい列は、空、デフォルト値、または式に基づいてポピュレートできる。
列分割（column split）	カラムを複数のカラムに分割。得られる各列は、元の列の一部を有する。
グループ分け（group by）	指定した列の固有値に基づいて行をグループ化
集約（aggregate）	行のグループ上で式を計算
ソート（sort）	指定した条件に基づいて行または列を並べ替え

【００２６】

データベースオペレーション履歴モジュール１１２はさらに、抽出されたデータに関するコンテキストデータを受信し、作成し、管理するように構成される。コンテキストデータは、テーブルまたはプロジェクトに対して実行されているデータベースオペレーションに関連して収集または生成されるテーブルおよび／またはプロジェクトに関する情報である。コンテキストデータは、プロジェクトメタデータ、テーブルメタデータ、列メタデータ、およびユーザメタデータを含む。コンテキストデータは、データベースオペレーション履歴ストア１２０に格納されてもよい。

【００２７】

プロジェクトメタデータフィールドは、以下を含む。

【表２】

表２：プロジェクトメタデータフィールド

メタデータフィールド	説明	変数タイプ
project_id	プロジェクト ID	英数字/数字の固有 ID
project_name	プロジェクト名	テキスト文字列
num_worksheets	プロジェクト内のテーブル数	整数
num_join_worksheets	プロジェクト内でジョインされたテーブル数	整数
num_union_worksheets	プロジェクト内で結合されたテーブル数	整数
num_agg_worksheets	プロジェクト内で実行された集合オペレーションの数	整数

【００２８】

テーブルメタデータフィールドは以下を含む。

【表３】

表３：テーブルメタデータフィールド

メタデータフィールド	説明	変数タイプ
ws_id	テーブル ID	英数字/数字の固有 ID
ws_name	テーブル名	テキスト文字列
ws_type	ワークシートタイプ	値に基づくテーブルタイプ
ws_rows	テーブル内の行数	整数
ws_curr_size	テーブルサイズ（単位？）	整数
ws_unique_cols	テーブル内の固有列数	整数
ws_text_cols	テキストを含む列の数	整数
ws_date_cols	日付を含む列の数	整数
ws_numeric_cols	数値を含む列の数	整数（whole number）
ws_blank_cols	空白列の数	整数（whole number）
ws_hidden_cols	隠された列の数	整数（whole number）
ws_derived_cols	派生値を含む列の数	整数（whole number）
recipe	テーブル上で実行されたオペレーションのシーケンス	テキスト文字列のリスト

【００２９】

列メタデータフィールドフィールドは以下を含む。

【表４】

表４：列メタデータフィールド

メタデータフィールド	説明	変数タイプ
column_id	列 ID	英数字の固有 ID
column_name	列名	テキスト文字列
column_datatype	列内のデータタイプ	
column_nulls	列内のヌル値パーセンテージ	小数
column_unique	列内の固有値パーセンテージ	小数
column_trimmable	列内のトリミング可能な値のパーセンテージ	小数
column_outlier	列内の外れ値のパーセンテージ	小数
column_pattern	カラム値のパターン	小数
column_domain	列のドメイン	テキスト文字列
column_selection	選択された列の値	テキスト文字列
column_maxvalue	列内の最大値	整数
column_minvalue	列内の最小値	整数

【００３０】

一実施形態では、コンテキストデータは、プロジェクトメタデータ、テーブルメタデータ、列メタデータ、ユーザメタデータ、および操作を含むログファイルに含まれる。ログファイル内のログエントリは、テーブルまたはプロジェクト上で実行されているデータベースオペレーションに応答して生成され、ＪａｖａＳｃｒｉｐｔ Ｏｂｊｅｃｔ Ｎｏｔａｔｉｏｎ（ＪＳＯＮ）で表すことができる。ログエントリは、コンテキストおよび操作履歴データを以下の形式で表現する。

{<user metadata><project metadata><worksheets metadata><column metadata><operation specifics>}

【００３１】

ログ・エントリは、データベースオペレーション履歴記憶装置１２０に記憶することができる。ログエントリの例を以下に示す：

【００３２】

ログエントリ例１

【数１】


```
{
  "type": "com.informatica.dataprep.suggestion.logger.ContextCollectorImpl",
  "data": {
    "userLogger": {
      "type": "com.informatica.dataprep.suggestion.logger.UserContextImpl",
      "data": {
        "user_id": 197,
        "licence_plan": "NA"
      }
    },
    "projectLogger": {
      "type": "com.informatica.dataprep.suggestion.logger.ProjectContextImpl",
      "data": {
        "project_id": 2312,
        "project_name": "test-log",
        "num_worksheets": 1,
        "num_join_worksheets": 0,
        "num_union_worksheets": 0,
        "num_agg_worksheets": 0
      }
    },
    "sheetLogger": [
      {
        "type": "com.informatica.dataprep.suggestion.logger.SheetContextImpl",
        "data": {
          "ws_id": 2313,
          "ws_name": "dp_user_session.csv",
          "ws_type": "NORMAL",
          "ws_rows": 31275,
          "ws_curr_size": 6,
          "ws_unique_cols": 3,
          "ws_text_cols": 3,
          "ws_date_cols": 0,
          "ws_numeric_cols": 3,
          "ws_blank_cols": 0,
          "ws_hidden_cols": 0,
          "ws_derived_cols": 0,
          "recipe": "deleteHeaderRows;"
        }
      }
    ],
    "columnLogger": [
      {
        "type":
"com.informatica.dataprep.suggestion.logger.ColumnContextImpl",
        "data": {
          "column_id": 2327,
          "column_name": "D",
          "column_datatype": "Integer",
          "column_nulls": 0.0,
          "column_unique": 99.81,
          "column_trimmable": 0.0,
          "column_outlier": 36.17585931254996,
          "column_pattern": "<NUMBER>",
          "column_domain": "None",
          "column_selection": "None",
          "column_maxvalue": "1427703590101",
          "column_minvalue": "1403021779000"
        }
      }
    ],
```

【数２】

```
"operationLogger": {
      "type":
"com.informatica.dataprep.suggestion.logger.OperationContextImpl",
      "data": {
        "operation": "expr:",
        "operation_description": "expr:(((C3/60)/60)/24000)+DATE(1970,1,1)"
      }
    },
    "timestamp": 1427863549753
  }
}
```

【００３３】

ログエントリ例１では、ユーザメタデータが「ｕｓｅｒＬｏｇｇｅｒ」セクションに含まれる。このセクションの「ｔｙｐｅ」サブセクションは、データがユーザメタデータ（「．．．ＵｓｅｒＣｏｎｔｅｘｔＩｍｐ」）であることを示す。このセクションの「ｄａｔａ」サブセクションは、データベースオペレーションが実行されたときにデータ解析アプリケーション１２５のユーザを一意に識別するユーザ識別子値（「ｕｓｅｒ＿ｉｄ」：１９７）を含む。

【００３４】

プロジェクトメタデータは「ｐｒｏｊｅｃｔＬｏｇｇｅｒ」セクションに含まれる。このセクションの「ｔｙｐｅ」サブセクションは、データがプロジェクトコンテキストデータ（「．．．ＰｒｏｊｅｃｔＣｏｎｔｅｘｔＩｍｐ」）であることを示す。「ｄａｔａ」サブセクションは、データベースオペレーションが実行されたプロジェクトの特徴を含む。この特徴は、プロジェクト識別子（「ｐｒｏｊｅｃｔ＿ｉｄ」：２３１２）、プロジェクト名（「ｐｒｏｊｅｃｔ＿ｎａｍｅ」：「ｔｅｓｔ－ｌｏｇ」）、プロジェクト内のテーブルの数（「ｎｕｍ＿ｗｏｒｋｓｈｅｅｔｓ」：１）、ジョインされたワークシートの数（「ｎｕｍ＿ｊｏｉｎ＿ｗｏｒｋｓｈｅｅｔｓＬ」：０）、接合されたワークシートの数（「ｎｕｍ＿ｕｎｉｏｎ＿ｗｏｒｋｓｈｅｅｔｓ」：０）、および集合ワークシートの数（「ｎｕｍ＿ａｇｇ＿ｗｏｒｋｓｈｅｅｔｓ」：０）を含む。

【００３５】

テーブルメタデータは、「ｓｈｅｅｔＬｏｇｇｅｒ」セクションに含まれる。このセクションの「ｔｙｐｅ」サブセクションは、データがテーブルメタデータ（「ＳｈｅｅｔＣｏｎｔｅｘｔＩｍｐｌ」）であることを示す。「ｄａｔａ」サブセクションは、データベースオペレーションが実行されたテーブルの特性を含む。この特性は、テーブル識別子（「ｗｓ＿ｉｄ」：２３１３）、テーブル名（「ｗｓ＿ｎａｍｅ」：「ｄｐ＿ｕｓｅｒ＿ｓｅｓｓｉｏｎ．ｃｓｖ」）、テーブルタイプ（「ｗｓ＿ｔｙｐｅ」：「ＮＯＲＭＡＬ」）、テーブル内の行数（「ｗｓ＿ｒｏｗｓ」：３１２７５）、テーブルサイズ（「ｗｓ＿ｃｕｒｒ＿ｓｉｚｅ」：６）、テーブル内の固有の列数（「ｗｓ＿ｕｎｉｑｕｅ＿ｃｏｌｓ」：３）、テーブル内のテキストの列数（「ｗｓ＿ｔｅｘｔ＿ｃｏｌｓ」：３）、日付形式の列数（「ｗｓ＿ｄａｔｅ＿ｃｏｌｓ」：０）、数の列数（「ｗｓ＿ｎｕｍｅｒｉｃ＿ｃｏｌｓ」：３）、空白の列数（「ｗｓ＿ｂｌａｎｋ＿ｃｏｌｓ」：０）、隠れ列数（「ｗｓ＿ｈｉｄｄｅｎ＿ｃｏｌｓ」：０）、派生列数（「ｗｓ＿ｄｅｒｉｖｅｄ＿ｃｏｌｓ」：０）、およびテーブル上で実行される操作のリスト（「ｒｅｃｉｐｅ」：「ｄｅｌｅｔｅＨｅａｄｅｒＲｏｗｓ；」）を含む。

【００３６】

列メタデータは「ｃｏｌｕｍｎＬｏｇｇｅｒ」セクションに含まれる。このセクションの「ｔｙｐｅ」サブセクションは、データが列メタデータ（「．．．ＣｏｌｕｍｎＣｏｎｔｅｘｔＩｍｐｌ」）であることを示す。「ｄａｔａ」サブセクションは、データベースオペレーションが実行された列の特性を含む。この特性は、列識別子（「ｃｏｌｕｍｎ＿ｉｄ」：２３２７）、列名（「ｃｏｌｕｍｎ＿ｎａｍｅ」：「Ｄ」）、列データ型（「ｃ

ｏｌｕｍｎ＿ｄａｔａｔｙｐｅ」：「Ｉｎｔｅｇｅｒ」）、列内のヌル値のパーセンテージ（「ｃｏｌｕｍｎ＿ｎｕｌｌｓ」：０．０）、列内の一意の値のパーセンテージ（「ｃｏｌｕｍｎ＿ｕｎｉｑｕｅ」：９９．８１）、トリミングのパーセンテージ、列内の可能な値（「ｃｏｌｕｍｎ＿ｔｒｉｍｍａｂｌｅ」：０．０）、列内の外れ値のパーセンテージ（「ｃｏｌｕｍｎ＿ｏｕｔｌｉｅｒ」：３６．１７５８５９３１２５５４９９６）、列の値のパターン（「ｃｏｌｕｍｎ＿ｐａｔｔｅｒｎ」：「＜ＮＵＭＢＥＲ＞」）、列のドメイン（「ｃｏｌｕｍｎ＿ｄｏｍａｉｎ」：「Ｎｏｎｅ」）、列の選択された領域（「ｃｏｌｕｍｎ＿ｓｅｌｅｃｔｉｏｎ」：「Ｎｏｎｅ」）、列の最大値（「ｃｏｌｕｍｎ＿ｍａｘｖａｌｕｅ」：「１４２７７０３５９０１０１」）、および列の最小値（「ｃｏｌｕｍｎ＿ｍｉｎｖａｌｕｅ」：「１４０３０２１７７７９０００」）を含む。

【００３７】

データベースオペレーション履歴データは、「ｏｐｅｒａｔｉｏｎＬｏｇｇｅｒ」セクションに含まれる。このセクションの「ｔｙｐｅ」サブセクションは、データがオペレーション履歴データ（「ＯｐｅｒａｔｉｏｎＣｏｎｔｅｘｔＩｍｐｌ」）であることを示す。このセクションの「ｄａｔａ」サブセクションは、どのデータベースオペレーションが実行されたかを識別するオペレーション識別子（「ｏｐｅｒａｔｉｏｎ」：「ｅｘｐｒ：」）と、データベースオペレーションに含まれたオペランドを示すオペレーション記述（「ｏｐｅｒａｔｉｏｎ＿ｄｅｓｃｒｉｐｔｉｏｎ」：「ｅｘｐｒ（（（Ｃ３／６０）／６０）／２４０００）＋ＤＡＴＥ（１９７０，１，１）」）とを含む。この例では、オペレーションが、タイムスタンプミリ秒を日数に変換するために使用され、それらを日付１／１／１９７０に追加して、タイムスタンプの日付を取得する。

【００３８】

ログエントリ例２

【数３】


```
{
  "type": "com.informatica.dataprep.suggestion.logger.ContextCollectorImpl",
  "data": {
    "userLogger": {
      "type": "com.informatica.dataprep.suggestion.logger.UserContextImpl",
      "data": {
        "user_id": 352,
        "licence_plan": "NA"
      }
    },
    "projectLogger": {
      "type": "com.informatica.dataprep.suggestion.logger.ProjectContextImpl",
      "data": {
        "project_id": 688,
        "project_name": "DP Tables",
        "num_worksheets": 2,
        "num_join_worksheets": 0,
        "num_union_worksheets": 0,
        "num_agg_worksheets": 0
      }
    },
    "sheetLogger": [
      {
        "type": "com.informatica.dataprep.suggestion.logger.SheetContextImpl",
        "data": {
          "ws_id": 762,
          "ws_name": "dp_user.csv",
          "ws_type": "NORMAL",
          "ws_rows": 5420,
          "ws_curr_size": 7,
          "ws_unique_cols": 2,
          "ws_text_cols": 6,
          "ws_date_cols": 0,
          "ws_numeric_cols": 1,
          "ws_blank_cols": 0,
          "ws_hidden_cols": 0,
          "ws_derived_cols": 0,
          "recipe": "deleteHeaderRows;upper;"
        }
      },
      {
        "type": "com.informatica.dataprep.suggestion.logger.SheetContextImpl",
        "data": {
          "ws_id": 689,
          "ws_name": "dp_user_session.csv",
          "ws_type": "NORMAL",
```

【数４】


```
            "ws_rows": 31275,
            "ws_curr_size": 8,
            "ws_unique_cols": 3,
            "ws_text_cols": 3,
            "ws_date_cols": 1,
            "ws_numeric_cols": 3,
            "ws_blank_cols": 1,
            "ws_hidden_cols": 0,
            "ws_derived_cols": 0,
            "recipe": "deleteHeaderRows;expr
(((C3/60)/60)/24000)+DATE(1970,1,1);expr;"
          }
        }
      ],
      "columnLogger": [
        {
          "type":
"com.informatica.dataprep.suggestion.logger.ColumnContextImpl",
          "data": {
            "column_id": 802,
            "column_name": "id",
            "column_datatype": "Integer",
            "column_nulls": 0.0,
            "column_unique": 100.0,
            "column_trimmable": 0.0,
            "column_outlier": 35.283262594574644,
            "column_pattern": "<NUMBER>",
            "column_domain": "None",
            "column_selection": "None",
            "column_maxvalue": "5947716",
            "column_minvalue": "22"
          }
        },
        {
          "type":
"com.informatica.dataprep.suggestion.logger.ColumnContextImpl",
          "data": {
            "column_id": 794,
            "column_name": "USER_ID",
            "column_datatype": "Integer",
            "column_nulls": 0.0,
            "column_unique": 11.55,
            "column_trimmable": 0.0,
            "column_outlier": 11.210590266675194,
            "column_pattern": "<NUMBER>",
            "column_domain": "None",
            "column_selection": "None",
            "column_maxvalue": "5947257",
            "column_minvalue": "22"
          }
        }
      ],
      "operationLogger": {
        "type":
"com.informatica.dataprep.suggestion.logger.OperationContextImpl",
        "data": {
          "operation": "Join",
          "operation_description": "Join:762 FULL_OUTER 689"
        }
      },
      "timestamp": 1427888644562
    }
  }
```


【００３９】

ログエントリ１の例は、１つのテーブルに対して実行されたデータベースオペレーションに対応する。ログエントリ２の例は、２つのテーブルに対して実行されたデータベースオペレーションに対応する。２つのテーブルに対して実行されるデータベースオペレーションは、２つのテーブルからの列を結合するジョイン（ｊｏｉｎ）および結合（ｕｎｉｏ

ｎ）操作を含む。ログエントリ２の例では、ＳｈｅｅｔＬｏｇｇｅｒセクションで指定されているように、テーブルＩＤ７６２および６８９を有するテーブルに対して完全外部ジョイン操作が実行された。ログエントリ２の例は、２組のテーブルデータおよび２組の列データを有し、各組は、ジョイン操作が実行された２つのテーブルのうちの１つに対応する。

【００４０】

データ解析サーバ１０４のユーザモジュール１１５は、ユーザがデータ解析サーバ１０４とのアカウントを管理することを可能にする。ユーザモジュール１１５はさらに、データ解析アプリケーション１２５に関連するユーザ活動に対応するユーザ情報を受信し、記憶する。ユーザ情報はユーザの好み、ユーザに関連するコンピューティングデバイスに関する情報、様々なグループ（例えば、企業（ｅｎｔｅｒｐｒｉｓｅ）、組織（ｏｒｇａｎｉｚａｔｉｏｎ）など）とのユーザの関連、およびトレーニングユーザおよび／またはガイド付きユーザとしてのユーザのステータスを含み得る。トレーニングユーザはデータ解析アプリケーション１２５のユーザであり、そのデータベースオペレーションは、ガイド付きユーザにデータベースオペレーションを推薦するための予測モデルをトレーニングするために使用される。ガイド付きユーザは、トレーニングされた予測モデルからデータベースオペレーションの推奨を受信するデータ解析アプリケーション１２５のユーザである。ガイド付きユーザの１つ以上のセットはトレーニングユーザに関連するデータを使用してガイド付きユーザの推薦が生成されるように、トレーニングユーザの１つ以上のセットに関連付けられてもよい。

【００４１】

ガイド付きユーザおよび／またはトレーニングユーザとしてのユーザのステータス、ならびにガイド付きユーザのセットとトレーニングユーザとの間の関連付けは、システム管理者、他のユーザによって指定されてもよいし、自動的に指定されてもよい。例えば、グループ（例えば、組織または企業）は、データ解析アプリケーション１２５の上級ユーザをトレーニングユーザとして、経験の少ないユーザをガイド付きユーザとして指定することができる。トレーニングユーザのセットはまた、地理的領域またはデータ解析アプリケーション１２５による熟練度の尺度などのユーザ特性に基づいて、ユーザモジュール１１５によって自動的に決定されてもよい。ガイド付きユーザは、トレーニングユーザに関連付けられたトレーニングデータがガイド付きユーザに対する推薦を生成するために使用されるように、トレーニングユーザに関連付けられてもよい。結果として、トレーニングユーザの知識および経験は、データ解析サーバ１０４によって活用されて、ガイド付きユーザに有用な推薦を提供し得る。グループからのトレーニングユーザのセットを同じグループからのガイド付きユーザのセットに関連付けることにより、システムはユーザに、そのグループに特に関連する推薦を提供することができ、グループ内のユーザがグループ全体の一貫性を維持し、独自の情報（例えば、方程式、関数、およびデータ）を保護しながら、生産性を高めることができるようになる。

【００４２】

一実施形態では、トレーニングユーザおよびガイド付きユーザの複数のセットが存在する。特定のユーザは同時にトレーニングユーザおよびガイド付きユーザとすることができ、複数の組のトレーニングユーザおよび／またはガイド付きユーザに属することができる。ユーザはあるタイプのプロジェクト（例えば、会計）に関してはトレーニングユーザであってもよいが、別のタイプのプロジェクト（例えば、マーケティング）に関してはガイド付きユーザであってもよい。トレーニングユーザおよび／またはガイド付きユーザとしてのユーザのステータス、ならびにトレーニングユーザとガイド付きユーザとの間の任意の関連付けは、ユーザデータストア１１７に格納され得る。ユーザモジュール１１５は、特定のプロジェクトについて、ユーザのステータスを、ガイド付きユーザまたはトレーニングユーザのいずれかとして決定することができる。ユーザがガイド付きユーザである場合、ユーザモジュール１１５はさらに、推薦を生成するためにトレーニングユーザのどのセットが使用されるべきかを決定することができる。

【００４３】
データベースオペレーション推薦モジュール１１４は、コンテキストデータおよびデータベースオペレーション履歴データに基づいて、ユーザに対して推薦されるデータベースオペレーションを決定する。データベースオペレーション推薦モジュール１１４は、予測モデルに基づいてデータベースオペレーションを推薦する。データベースオペレーションは、予測モデルによっても決定されるオペランドを含む。オペランドは、関数入力などのデータベースオペレーションのための入力またはパラメータである。様々な実施形態では、予測モデルがデータベース動作履歴データおよびコンテキストデータを使用することによってトレーニングされ得る機械学習アルゴリズムである。ロジスティック回帰、ニューラルネットワーク、決定木モデル、およびサポートベクトルマシンモデルを含む、様々な予測モデルが当技術分野で周知である。モデルは入力の特定のセット（例えば、コンテキストデータ）が与えられると、特定のデータベースオペレーションが適切である確率を予測し、可能性のあるデータベースオペレーションのうちの１つまたは複数、および任意選択で、推奨された操作に対応するオペランドを推奨する。予測モデルは、データベース動作履歴データおよびコンテキストデータを使用してトレーニングされる機械学習アルゴリズムとすることができる。一実施形態では、多項ロジスティック分類器または他の適切な汎用機械学習技法などの識別モデルが使用される。方程式、パラメータ、および他のモデル特性は、データベースオペレーション推奨ストア１２１に格納され得る。データベースオペレーション推奨を生成するための３つのモデル例について、図２を参照して以下に説明する。

【００４４】
図２は、一実施形態によるデータベースオペレーション推薦モジュール１１４のより詳細な図を示す。モデル構築モジュール２０５は予測モデルを構築し、モデルトレーニングモジュール２１０はトレーニングユーザからのトレーニングデータを使用して予測モデルをトレーニングし、推薦生成モジュール２２０は、トレーニングされた予測モデルを使用して、ガイド付きユーザに対する推薦のためのデータベース動作を決定する。一実施形態では、モデルが多項ロジスティック分類子を使用する。ログエントリからプロファイルされたメタデータフィールドによって表されるような特定のコンテキストデータが与えられると、多項式ロジスティック分類子を使用するモデルは、それぞれの確率を有するデータベース演算のリストを生成する。モデルは、トレーニングデータを用いてトレーニングされる。一実施形態では、トレーニングデータがトレーニングユーザのセットに関する格納されたデータベースオペレーション履歴データおよびコンテキストデータを含む。この実施形態では、モデルトレーニングモジュール２１０が例えばユーザデータストア１１７からモデルのトレーニングユーザを決定し、データベース動作履歴ストア１２０からトレーニングデータを取り出す。

【００４５】
モデル構築モジュール２０５は予測モデルを構築し、この機能を実行するための１つの手段である。多項ロジスティック分類器は、所与の情報に基づいて事象が発生する確率の推定値を提供する。多項ロジスティック分類器は、以下の形式をとる：

【００４６】
【数５】

$$P(c|d) = \frac{\exp(\sum_i(\lambda_{i,c}F_{i,c}(d,c)))}{\sum_{c'}\exp(\sum_i(\lambda_{i,c'}F_{i,c'}(d,c')))}$$

【００４７】
ここで、Ｐ（ｃ｜ｄ）は、特徴Ｆによって特徴付けられる条件ｄが与えられた場合に生じるクラスｃによって特徴付けられる事象の確率の推定値である。クラスｃは、演算またはオペランドのいずれかである特定の予測モデルの出力に対応し、特徴Ｆは、関連するコ

ンテキストデータに対応する。F_i（ｄ，ｃ）は特徴ｉの観測の尺度であり、Ｆ値が高いほど、特徴の存在の相対的な尺度が高いことを示す。$\lambda_{i,c}$は、クラスｃに対応する特徴ｉの特徴重みである。特定の特徴に対する高い$\lambda_{i,c}$は、Ｆ値がクラスｃに対する強力な指標であることを示す。特徴は、異なるクラスｃに対して異なるＦ値またはλ値を有することができる。Ｐ（ｃ｜ｄ）によって表される確率は、クラスｃについて、クラスの全ての特徴にわたる観測の尺度と特徴の重みとの積の合計の指数（ｅｘｐｏｎｅｎｔｉａｌ）を決定し、その値を全てのクラスにわたる同じ値の合計で割ることによって、計算される。

【００４８】

一実施形態では、モデル構築モジュール２０５が３つのモデルを構築する：演算モデル（ＯＰモデル）、オペランドモデル（ＯＰＤモデル）、列演算モデル（ＯＰＣモデル）である。３つのモデルの各々は、トレーニングユーザーからのトレーニングデータを使用してモデルトレーニングモジュール２１０によってトレーニングされる。３つのモデルの各々は、推薦生成モジュール２２０によって使用されて、推薦されたデータベースオペレーションおよび／またはオペランドと、コンテキストデータに基づく関連する相対確率とのリストを生成する。

【００４９】

ＯＰモデルは、単一テーブルデータベースオペレーションの推奨データベースオペレーションのリストおよび関連する確率を生成する。ＯＰモデルの機能は、列メタデータフィールドである。

【００５０】

ＯＰＤモデルは、単一テーブルデータベースオペレーションのための推奨データベースオペレーションのためのオペランドのリストおよび関連する確率を生成する。ＯＰＤモデルの特徴は、列メタデータフィールドおよびデータベースオペレーションである。一実施形態では、ＯＰＤモデルがＯＰモデルと併せて使用され、ＯＰモデルによって決定されたデータベースオペレーションのためのオペランドを決定する。ＯＰＤモデルは、ＯＰＤモデルによって決定された推奨オペランドが決定された演算に対応するように、ＯＰモデルによって決定されたデータベース演算を入力として取り込む。

【００５１】

ＯＰＣモデルは、２テーブルデータベースオペレーションのための推奨データベースオペレーションのリストおよび関連する確率を生成する。ＯＰＣモデルの特徴は、２つのテーブルの各々および２つの列の各々についてのメタデータである。

【表５】

表５：モデルクラスと特徴

モデル	クラス	特徴
OP	データベースオペレーション	列メタデータ
OPD	オペランド	列メタデータ、データベースオペレーション
OPC	ジョイン、結合	テーブルメタデータ、列メタデータ

【００５２】

各モデルについて、モデルトレーニングモジュール２１０は、多項ロジスティック分類器に含める特徴としてどのコンテキストデータフィールドが選択されるかを決定する。モデルトレーニングモジュール２１０はさらに、選択された各特徴に対する特徴重みを決定する。すべてのメタデータフィールドが演算および／またはオペランドを予測するわけではないので、すべてのメタデータフィールドがモデルの特徴として使用されるわけではな

い。一実施形態では、モデルトレーニングモジュール２１０が複数のデータベース動作履歴エントリにわたって、取られる特定のデータベース動作または使用されるオペランドを予測するモデル特徴として使用するコンテキストデータフィールドを選択する。モデルトレーニングモジュール２１０は、各コンテキストデータフィールドについて予測性の尺度を計算し、予測性の尺度は例えば、情報利得であってもよい。各クラスについて、モデルトレーニングモジュール２１０は、格納されたコンテキストデータに基づいて、可能な特徴のリスト内の各特徴についての情報利得を計算する。モデルトレーニングモジュール２１０は、閾値情報ゲイン値を超える特徴を選択し、モデルに含める。所与のクラスについて、特徴に関する情報利得は、以下の式によって計算することができる。

ＩＧ（Ｃ｜Ｆ）＝Ｅｎｔｒｏｐｙ（Ｃ）－Ｅｎｔｒｏｐｙ（Ｃ｜Ｆ）

ここで、ＩＧ（Ｃ｜Ｆ）は情報利得であり、Ｅｎｔｒｏｐｙ（Ｃ）はクラスＣのエントロピーであり、Ｅｎｔｒｏｐｙ（Ｃ｜Ｆ）は特徴の存在を仮定したクラスＣの条件付きエントロピーである。

【００５３】

一実施形態では、モデルトレーニングモジュール２１０が情報利得を計算する前に、コンテキストデータを前処理する。一実施形態では、モデルトレーニングモジュール２１０がコンテキストデータを再サンプリングして、各クラスにわたるデータエントリの分布をより均一にし、その結果、より少ない頻度のデータベースオペレーションがモデルにおいて過少に表されないようにする。再サンプリング技術は、アンダーサンプリング法、オーバーサンプリング法、またはハイブリッド法を含むことができる。一実施形態では、リサンプリングがＳＭＯＴＥ（Ｓｙｎｔｈｅｔｉｃ　Ｍｉｎｏｒｉｔｙ　Ｏｖｅｒｓａｍｐｌｉｎｇ　Ｔｅｃｈｎｉｑｕｅ）を用いて実行される。様々な実施形態では、すべてのデータを数値表現に変換すること、データの正規化、および数値の２進数への量子化など、他の前処理ステップがコンテキストデータに対して実行される。

【００５４】

図３は、ＯＰモデルなどの予測モデルをトレーニングする際に使用するための特徴およびクラスを示すデータエントリの例示的なテーブルである。図３の例では、列３０１～３０８に示される特徴が、データベースオペレーションが実行された列に対応する列メタデータエントリの選択されたセットである。

【００５５】

列３０１は、表４で識別される「ｃｏｌｕｍｎ＿ｉｄ」メタデータフィールドからの値を含む。

【００５６】

列３０２は、表４で識別される「ｃｏｌｕｍｎ＿ｔｙｐｅ」メタデータフィールドからの値を含む。

【００５７】

列３０３は、表４で識別される「ｃｏｌｕｍｎ＿ｎｕｌｌｓ」メタデータフィールドからの値を含む。

【００５８】

列３０４は、表４で識別される「ｃｏｌｕｍｎ＿ｕｎｉｑｕｅ」メタデータフィールドからの値を含む。

【００５９】

列３０５は、表４で識別される「Ｃｏｌｕｍｎ＿ｐａｔｔｅｒｎ」メタデータフィールドからの値を含む。

【００６０】

列３０６は、表４で識別される「ｃｏｌｕｍｎ＿ｄｏｍａｉｎ」メタデータフィールドからの値を含む。

【００６１】

列３０７は、表４で識別される「ｃｏｌｕｍｎ＿ｍａｘｖａｌｕｅ」メタデータフィールドからの値を含む。

【００６２】
列３０８は、表４で識別される「ｃｏｌｕｍｎ＿ｍｉｎｖａｌｕｅ」メタデータフィールドからの値を含む。
【００６３】
ここでのモデルのクラスは列３１０に示される表１に識別されるように、列上で実行されたデータベースオペレーションの名前である。表５に示されるように、これらの特定の例示的な特徴およびクラスは、ＯＰモデルをトレーニングするために使用される。図３の例は１４個のデータエントリを示すが、実際には上述した予測モデルが数百、数千、数百万またはそれ以上のデータエントリを用いてトレーニングすることができる。様々な実施形態において、データエントリを構成するコンテキストデータおよびデータベースオペレーション履歴データの断片は、図２に関して上述したように、モデルトレーニングモジュール２１０によってログエントリから選択される。データエントリは、データベースオペレーション推薦ストア１２１に格納されてもよい。
【００６４】
推薦生成モジュール２２０は、トレーニングされた予測モデルを使用して、ガイド付きユーザに対する推薦のためのデータベースオペレーションおよび／またはオペランドのリストを、それぞれの相対確率と共に決定する。推薦生成モジュール２２０は、例えばログファイルの形式でコンテキストデータを受信する。推薦生成モジュール２２０は関連コンテキストデータを、予測モデルに入力され得るフォーマットでキャプチャするように、ログファイルをプロファイルする。推薦生成モジュール２２０は推薦を生成するために、適切な予測モデルにコンテキストデータを入力する。様々な実施形態では、使用される予測モデルが単一テーブル推奨の場合にはＯＰモデルおよびＯＰＤモデルであり、マルチテーブル推奨の場合にはＯＰＣモデルである。推薦生成モジュール２２０は、様々なイベントの発生時に、定期的な間隔で、または任意の他の適切な時間に、推薦を生成することができる。一実施形態では、推薦生成モジュール２２０がデータ解析アプリケーション１２５のユーザインタフェースにおける列の選択を検出し、それに応答してその列に対する推薦を生成するプログラムコードを実行する。このような推薦を生成するためのプロセスは、図５に関して以下に説明される。
【００６５】
推薦生成モジュール２２０は、生成された１つまたは複数のリストから１つまたは複数の推薦データベースオペレーションおよび／またはオペランドを選択する。一実施形態では、推薦生成モジュール２２０が予測モデルによって計算されるように、最も高い相対確率を有する推薦を選択する。例えば、選択された列に対する単一シート推薦に対して、推薦生成モジュール２２０は、ＯＰモデルによって決定される３つの最も確からしいデータベースオペレーションと、ＯＰＤモデルによって決定される各オペレーションに対する１つの最も確からしいオペランドとを選択することができる。
【００６６】
推薦生成モジュール２２０はユーザに表示するために、データ解析アプリケーション１２５に推薦を提供する。一実施形態では、推奨が動作のテキスト記述として提供される。各データベースオペレーションのテキスト記述は、データベースオペレーション推薦ストア１２１に格納することができる。推薦生成モジュール２２０はユーザに表示するためにデータ解析アプリケーション１２５に提供するために、推薦されたデータベースオペレーションのためのテキスト記述を取り出してもよい。
【００６７】
図４は、一実施形態による、データ解析アプリケーションにおいてデータを閲覧および操作するためのユーザインタフェース４００の一例を示す。例示的なユーザインタフェースは、データセクション４１０、情報セクション４１５、およびコントロール４１７を含む。
【００６８】
データセクション４１０は、閲覧および操作のためのテーブルを表示する。データセク

ション４１０は１つ以上のデータソース（例えば、１０２）から抽出されたデータでポピュレート（ｐｏｐｕｌａｔｅｄ）される。この例では、２つのテーブルタブ４０５が示され、「ＭＤＭ顧客データ（ＭＤＭ　Ｃｕｓｔｏｍｅｒ　Ｄａｔａ）」と題するテーブルがデータセクション４１０に表示される。ユーザは、テーブルタブ４０５を使用してプロジェクト内の他のテーブルにナビゲートすることができる。図４の例では、列「ｆｉｒｓｔ＿ｎａｍｅ」４０７が選択される。

【００６９】

情報セクション４１５は、テーブルおよび選択されたデータに関するプロファイル情報を表示する。情報セクション４１５において、オーバービューカード４２０は選択された列（ｆｉｒｓｔ　ｎａｍｅ）の情報オーバービュー（例えば、タイプ、固有値のパーセンテージ、ブランク値のパーセンテージ、列内の名前の最小長、列内の名前の最大長、およびドメインの数）を提供する。ドメインカード４２５は、テーブル４０５内のすべてのドメインに関する情報、およびどのくらいの行が各ドメインに対応するかについての情報を含む。値頻度カード４３０は、選択されたファーストネーム列４０７における種々の名前の値の頻度、並びに名前の各時間がどのように発生するかをリストする。

【００７０】

提案カード４３５は、データベースオペレーション推奨モジュール１１４によって決定された推奨データベースオペレーションを実行するための提案をユーザに提供する。図示の例では、提案されたデータベース動作がファーストネーム（ｆｉｒｓｔ　ｎａｍｅ）として検証される。システムは、インタフェースのユーザにこれらのインテリジェントな提案を提供するのを助けるために、上述のデータプロファイリングを使用する。提案カード４３５については、図５および図６に関して以下でより詳細に説明する。

【００７１】

コントロール３１７は、ユーザが表示されたデータおよびテーブルを操作することを可能にし、データおよびテーブルに対してデータベースオペレーションを実行することを含む。データおよびテーブルは、データエントリとの対話（セル内容の編集、セルの右クリック、方程式の挿入など）または提案カード３３５などの情報セクション内の要素との対話などの他の方法で操作することもできる。

【００７２】

図５Ａは一実施形態による、データ解析アプリケーションのガイド付きユーザに対してデータベースオペレーションを決定し推奨するための予測モデルを構築し、トレーニングするための方法を示すフローチャートである。データ解析サーバ１０４は、データ解析アプリケーション１２５のユーザをトレーニングするためのコンテキストデータおよびデータベースオペレーション履歴データを維持する（５００）。データ解析サーバ１０４はある期間にわたって、データ解析アプリケーション１２５のインスタンスからのコンテキストデータおよびデータベースオペレーション履歴データを、例えば、図１に関して上述したようなログファイルとして受信および格納することによって、コンテキストデータおよびデータベースオペレーション履歴データを維持する。一実施形態では、データ解析アプリケーション１２５がデータベースオペレーションを検出すると、ログファイルをデータベースオペレーション履歴モジュール１１２に送信する。別の実施形態では、データベースオペレーション履歴モジュール１１２がデータ解析アプリケーション１２５を継続的に監視し、データベースオペレーションを検出すると、データベースオペレーション履歴データおよび対応するコンテキストデータを受信し、記憶する。

【００７３】

図１に関して上述したように、ガイド付きユーザおよび／またはトレーニングユーザとしてのユーザのステータス、ならびにガイド付きユーザとトレーニングユーザとの組の間の関連付けは、システム管理者、他のユーザによって、または自動的に指定することができる。

【００７４】

ステップ５０５および５１０では、ガイド付きユーザに推薦を提供する際に使用するた

めに、１つまたは複数の予測モデルが構築され、トレーニングされる。データベースオペレーション推奨モジュール１１４は、予測モデルを構築する（５０５）。予測モデルは、演算モデル（ＯＰ）、オペランドモデル（ＯＰＤ）、列演算モデル（ＯＰＣ）、またはそれらの任意の組合せとすることができる。予測モデルを構築することは、そのデータベースオペレーションがモデルのトレーニングデータとして使用されるトレーニングユーザを決定することを含む。予測モデルを構築することは、モデルクラスを決定することをさらに含む。例えば、予測モデルがＯＰモデルである場合、クラスはデータベースオペレーションである。予測モデルがＯＰＤモデルである場合、クラスはオペランドである。予測モデルがＯＰＣモデルである場合、クラスは結合および和演算、または定義された２テーブル演算である。予測モデルを構築するステップは上記の表５に関して説明したように、可能なモデル特徴を決定するステップをさらに含む。予測モデルを構築するステップは、データベース動作推薦ストア１２１からモデル方程式を検索するステップをさらに含む。ステップ５０５の終わりに、モデルはそのトレーニングされていない形式で存在する。図２に関して説明した方程式は各クラスについて組み立てられるが、特徴重みは未知であるか、またはデフォルト値に設定される。この形態では、モデルが決定されたトレーニングユーザに対応する適切なコンテキストデータを用いてトレーニングする準備ができている。

【００７５】

モデルトレーニングモジュール２１０は、決定されたトレーニングユーザからの維持されたデータベースオペレーション履歴データおよびコンテキストデータを使用して、モデルをトレーニングする（５１０）。モデルトレーニングモジュール２１０は、プロファイリングデータストア１１８およびデータベースオペレーション履歴ストア１２０から、トレーニングユーザに対応するデータベースオペレーション履歴データおよびトレーニングコンテキストデータを検索する。図２に関して上述したように、モデルトレーニングモジュール２１０は、どのコンテキストデータが特定のデータベースオペレーションまたはオペランドを予測するかを決定する。モデルトレーニングモジュール２１０は、図２に関して上述したように、各モデル特徴に対する特徴重みを決定する。特徴重みおよび他のパラメータは、データベースオペレーション推薦ストア１２１に格納され、必要に応じて使用のために取り出され得る。一実施形態では、モデルトレーニングモジュール２１０が図２に関して上述したように、モデルをトレーニングする前にコンテキストデータを前処理する。一旦、モデルがトレーニングされると、モデルは、特徴のセット（データ解析アプリケーションから受信したコンテキストデータ）に基づいてクラス（オペレーションまたはオペランド）の確率を決定するために使用され得る。

【００７６】

ステップ５０５および５１０は定期的な間隔で、継続的に、またはどれだけの新しいトレーニングデータが利用可能であるかなどの要因に応じて、行われ得る。ステップ５０５および５１０は、データベース動作推奨モジュール１１４によって生成される各予測モデルに対して繰り返されてもよい。図２に関して上述したように、ＯＰＤモデルはＯＰモデルによって決定されたデータベースオペレーションのためのオペランドを決定するために、ＯＰモデルと共に使用することができる。ＯＰＤモデルはＯＰＤモデルによって決定された推奨オペランドが決定された演算に対応するように、ＯＰモデルによって決定されたデータベース演算を入力として取り込むことができる。

【００７７】

図５Ｂは、一実施形態による、データ解析アプリケーションのガイド付きユーザにデータベースオペレーションを推薦するためにトレーニングされた予測モデルを使用する方法を示すフローチャートである。推薦生成モジュール２２０は、ガイド付きユーザのデータ解析アプリケーション１２５からアプリケーションコンテキストデータを受信する（５５０）。一実施形態では、アプリケーション・コンテキスト・データがデータ解析アプリケーション内に表示されたテーブル内で選択された列など、データ解析アプリケーション１２５との検出された対話に応答して受信される。データ解析アプリケーション１２５は対話を検出し、コンテキストデータを含むアプリケーションログエントリを作成し、アプリ

ケーションログエントリをデータ解析サーバ１０４に送信する。一実施形態では、推薦生成モジュール２２０が、アプリケーションログエントリをプロファイルして、トレーニングされた予測モデルへの入力として使用することができるフォーマットで、コンテキストデータを取り込む。

【００７８】

推薦生成モジュール２２０は、コンテキストデータに基づいて、推薦を生成するために使用する１つ以上のモデルを選択する（５５５）。例えば、コンテキストデータが、プロジェクトが１つのテーブルを有することを示す場合、推薦生成モジュール２２０は、ＯＰモデルおよびＯＰＤモデルを使用して、推薦を生成する。コンテキストデータが、プロジェクトが複数のテーブルを有することを示す場合、推薦生成モジュール２２０は、ＯＰモデル、ＯＰＤモデル、およびＯＰＣモデルを使用して、推薦を生成する。図２および図５Ａに関して上述したように、ＯＰＤモデルは、ＯＰモデルの出力を入力として使用して、ＯＰモデルによって決定された推奨演算のリストに対応するオペランドを決定することができる。

【００７９】

推薦生成モジュール２２０は、選択された予測モデルおよび受信されたコンテキストデータを使用して、ガイド付きユーザに推薦するためのデータベースオペレーションおよび／またはオペランドのリストを生成する（５６０）。様々な実施形態では、生成された推奨のリストがＯＰモデル、ＯＰＤモデル、およびＯＰＣモデル、ならびに他の予測モデルのうちの１つまたは複数によって決定される演算およびオペランドを含む。推薦生成モジュール２２０は、ステップ５５５で選択された各モデルを使用して、各モデルクラスに関連する確率を決定する。生成された推奨のリストは、決定された確率に基づいている。例えば、ＯＰモデルまたはＯＰＣモデルが使用される場合、推薦生成モジュール２２０はモデルによって決定されるような多数の最も確からしいデータベースオペレーションを選択し、案内されたユーザに推薦として提供する。ＯＰＤモデルも使用される場合、ＯＰモデルによって決定された選択されたデータベース演算は、選択されたデータベース演算のための最も可能性の高いオペランドの数を決定するためにＯＰＤモデルへの入力として使用される。

【００８０】

推薦生成モジュール２２０は、ガイド付きユーザに提示するために、推薦のリストをデータ解析アプリケーション１２５に送信する（５３５）。一実施形態では、各推奨データベースオペレーションがデータ解析アプリケーション１２５のデータベースオペレーションを一意に識別するオペレーション識別子を含む。別の実施形態では、各推奨データベースオペレーションがデータ解析アプリケーション１２５のユーザに提示するためのデータベースオペレーションのテキスト名または説明をさらに含む。データベースオペレーション、オペレーション識別子、ならびにテキスト名および説明は、データベースオペレーション推薦ストア１２１に格納され、推薦されたデータベースオペレーションをデータ解析アプリケーション１２５に送る前に、データベースオペレーション推薦モジュール１１４によって検索され得る。

【００８１】

図６は、一実施形態による、選択された列に応答して提供される推奨を備えた、図３の例示的なユーザインタフェースを示す。例示的なユーザインタフェースでは、例えばユーザ入力に応答して、列６５０が選択される。データ解析アプリケーション１２５は列選択を検出し、データ解析サーバ１０４に通知する。データ解析サーバ１０４は、列６５０の選択に応答して、データ解析アプリケーション１２５からコンテキストデータを受信する。一実施形態では、データベースオペレーション推薦モジュール１１４がユーザのステータスを、ユーザデータストア１１７からの特定のプロジェクトに対するガイド付きユーザと決定し、コンテキストデータをＯＰモデル（単一の列が選択されるため）およびＯＰＣモデルに渡す。ＯＰモデルは演算のリストを出力し、ＯＰＣモデルは、１つ以上のオペランドを出力する。データベースオペレーション推奨モジュール１１４は、推奨されるデー

タベースオペレーション、および適切な場合にはオペランドを決定し、その推奨をデータ解析アプリケーション１２５に送る。図６の例では、ユーザが異なる方法でフォーマットされた電話番号を含むように見える列を選択している。したがって、提供される２つの推奨は、ＯＰモデルによって決定された電話番号をフォーマットする動作と、ＯＰＤモデルによって決定された適用する特定の形式のフォーマットのオペランドとを含む。

【００８２】

図７は、データ解析アプリケーションにおいて、データ解析サーバから受信した推奨データベースオペレーションおよびオペランドを提示するための方法を示すフローチャートである。データ解析アプリケーション１２５は、データ解析サーバ１０４から推奨データベースオペレーションおよびオペランドを受信する（７００）。図５に関して上述したように、データベース動作は、データ解析アプリケーション１２５のユーザインタフェースに提示するためのテキスト名または記述を含むことができる。ＵＩモジュール１２２は、データ解析サーバ１０４によって提供されるテキスト名および説明を使用して、推奨データベースオペレーションおよびオペランドに対応するユーザインタフェース要素を生成する（７１０）。ＵＩモジュール１２２は、データ解析アプリケーション１２５のユーザインタフェースを介して、データ解析アプリケーションのユーザに１つ以上の推奨データベースオペレーションを提示する（７２０）。

【００８３】

図６に戻ると、提案カード４３５は推奨データベースオペレーションを含む。列６５０は、異なる方法でフォーマットされた電話番号を含む。提案カード４３５上の推奨６６０Ａ～Ｃは、セルまたは列内の電話番号をフォーマットすることを含む。推奨６５０ＡおよびＢは、共通のデータベースオペレーション（電話番号のフォーマット）を有するが、異なるオペランド（電話番号の出力フォーマット）を有する。データ解析アプリケーションのユーザは、推奨６６０Ａ～Ｃのうちの１つを選択して、データに対して指示されたデータベースオペレーションを実行することができる。

【００８４】

追加構成の考慮事項

本明細書で説明するシステムは、クラウドベースのコンピュータ実装を含む、単一のコンピュータまたはコンピュータのネットワークを使用して実装することができる。コンピュータは、好ましくは１つまたは複数の高性能ＣＰＵおよび１Ｇまたはそれ以上のメインメモリ、ならびに５００Ｇｂから２Ｔｂのコンピュータ可読永続ストレージを含み、ＬＩＮＵＸまたはその変形などのオペレーティングシステムを実行するサーバクラスコンピュータである。本明細書で説明するシステムの動作は、コンピュータストレージにインストールされ、本明細書で説明する機能を実行するために、そのようなサーバのプロセッサによって実行されるハードウェアおよびコンピュータプログラムの組み合わせによって制御することができる。システム１００はネットワークインターフェースおよびプロトコル、データ入力のための入力デバイス、ならびに表示、印刷、または他のデータの提示のための出力デバイスを含む、本明細書で説明される動作に必要な他のハードウェア要素を含むが、これらは実施形態の関連する詳細を不明瞭にすることを避けるために本明細書では示されない。

【００８５】

上記の説明のいくつかの部分は、アルゴリズムのプロセスまたは動作に関して実施形態を説明する。これらのアルゴリズムの説明および表現は、データ処理技術の当業者によって一般的に使用され、彼らの作業内容を他の当業者に効果的に伝える。これらの動作は機能的、計算的、または論理的に説明されているが、プロセッサまたは等価の電気回路、マイクロコードなどによって実行される命令を含むコンピュータプログラムによって実施されるものと理解される。さらに、一般性を損なうことなく、これらの機能的オペレーションの配置をモジュールと呼ぶことも便利である場合があることが判明している。説明された動作およびそれらの関連するモジュールは、ソフトウェア、ファームウェア、ハードウェア、またはそれらの任意の組合せで具現化されてもよい。

【００８６】
本明細書で使用されるように、用語「モジュール」は、指定された機能を提供するために利用されるコンピュータプログラムロジックを指す。したがって、モジュールは、ハードウェア、ファームウェア、および／またはソフトウェアで実装することができる。一実施形態では、プログラムモジュールが記憶装置に格納され、メモリにロードされ、プロセッサによって実行される。本明細書で説明される物理的構成要素の実施形態は、本明細書で説明されるもの以外の他のおよび／または異なるモジュールを含むことができる。さらに、他の実施形態では、モジュールに起因する機能が他のモジュールまたは異なるモジュールによって実行することができる。さらに、この説明は、明瞭さおよび便宜のために「モジュール」という用語を省略することがある。

【００８７】
本発明はまた、本明細書における動作を実行するための装置に関する。この装置は、必要な目的のために特別に構築されてもよく、またはコンピュータによってアクセスされ得るコンピュータ可読媒体上に格納されたコンピュータプログラムによって選択的に起動または再構成される汎用コンピュータを備えてもよい。そのようなコンピュータプログラムはフロッピー（登録商標）ディスク、光ディスク、ＣＤ－ＲＯＭ、磁気－光ディスク、読み取り専用メモリ（ＲＯＭ）、ランダムアクセスメモリ（ＲＡＭ）、ＥＰＲＯＭ、ＥＥＰＲＯＭ、磁気または光カード、特定用途向け集積回路（ＡＳＩＣ）、または電子命令を記憶するのに適した任意のタイプのコンピュータ可読記憶媒体を含む任意のタイプのディスクなどのコンピュータ可読記憶媒体に記憶することができ、それぞれがコンピュータシステムバスに結合されるが、これらに限定されない。さらに、本明細書で言及するコンピュータは単一のプロセッサを含むことができ、または計算能力を高めるために複数のプロセッサ設計を使用するアーキテクチャとすることができる。

【００８８】
本明細書で使用される「１つの実施形態」または「一実施形態」への言及は、実施形態に関連して説明された特定の要素、機能、構成、または特徴が少なくとも１つの実施形態に含まれることを手段する。明細書の様々な場所における「一実施形態では」という語句の出現は、必ずしもすべてが同じ実施形態を指すとは限らない。

【００８９】
本明細書において用いられるとき、「備える（ｃｏｍｐｒｉｓｅｓ）」、「備える（ｃｏｍｐｒｉｓｉｎｇ）」、「含む（ｉｎｃｌｕｄｅｓ）」、「含める（ｉｎｃｌｕｄｉｎｇ）」、「有する（ｈａｓ）」、「有する（ｈａｖｉｎｇ）」という用語またはそれらの任意の他の活用形は、非限定的な包含をカバーするものとする。例えば、一連の要素を含むプロセス、方法、物品、または装置は、それらの要素のみに必ずしも限定されず、特に明記されていないかあるいはかかるプロセス、方法、物品、または装置に固有の他の要素を含めてもよい。更に、明確に逆のことを表さない限り、「または」は包括的「ｏｒ」を指し、排他的「ｏｒ」を意味しない。例えば、条件ＡまたはＢは、以下のいずれか１つによって満たされる：Ａが真であり（または存在する）かつＢが偽である（または存在しない）、Ａが偽であり（または存在しない）かつＢが真である（または存在する）、ＡおよびＢの両方が真である（または存在する）。

【００９０】
さらに、「１つの（ａ）」または「１つの（ａｎ）」の使用は、本明細書の実施形態の要素および構成要素を説明するために使用される。これは、単に便宜上、かつ本開示の一般的な意味を与えるためになされる。本明細書は１つまたは少なくとも１つを含めるように読まれるべきであり、複数でないことを意図することが明白でない限り、単数形は複数形も含める。

【００９１】
本開示を読めば、当業者は、識別子空間にわたるエンティティの類似性を決定するためのシステムおよびプロセスのためのさらに追加の代替の構造および機能設計を理解するであろう。したがって、特定の実施形態および用途を図示し、説明したが、本発明は、本明

細書に開示される正確な構成および構成要素に限定されず、当業者には明らかな様々な修正、変更、および変形が添付の特許請求の範囲に定義される精神および範囲から逸脱することなく、本明細書に開示される方法および装置の構成、動作、および詳細において行われてもよいことを理解されたい。

【図１】

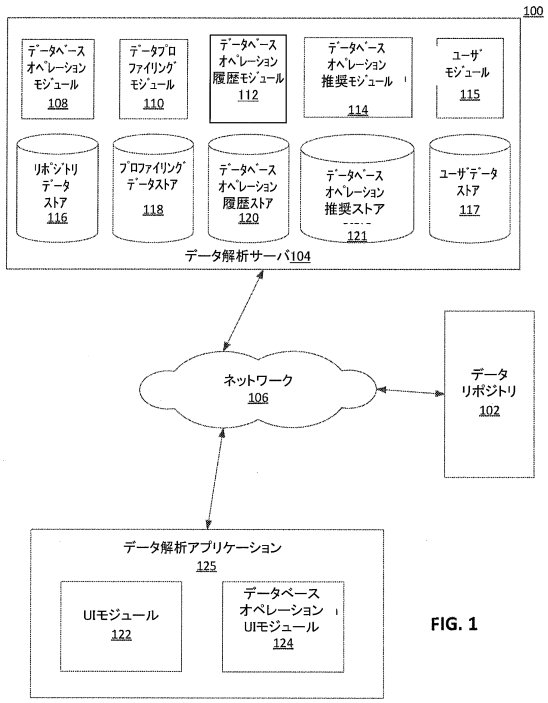


FIG. 1

【図２】

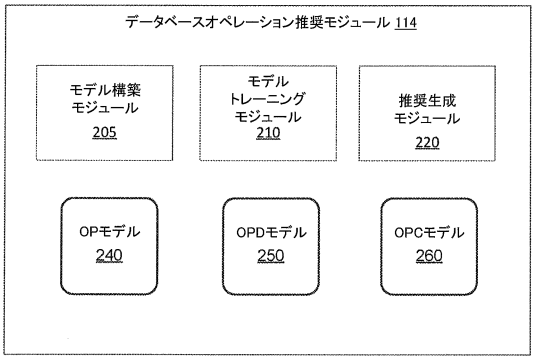
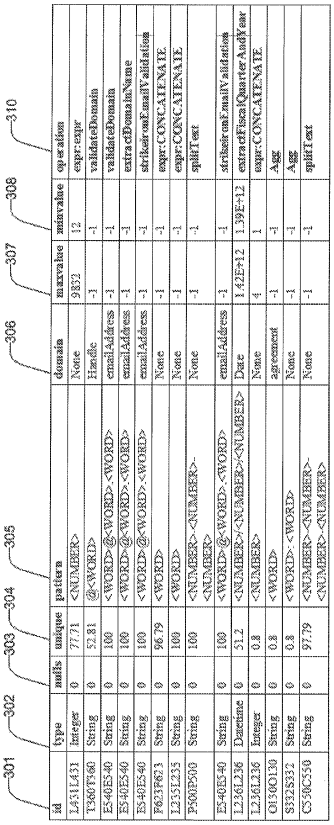


FIG. 2

【図３】



id	type	nulls	unique	pattern	domain	maxvalue	minvalue	operation
L3511431	Integer	0	77.71	<NUMBER>	None	9352	12	expr:expr
T360T960	String	0	52.81	@<WORD>	Handle	-1	-1	validateDomain
E540ES40	String	0	100	<WORD>@<WORD>.<WORD>	emailAddress	-1	-1	validateDomain
E540ES40	String	0	100	<WORD>@<WORD>.<WORD>	emailAddress	-1	-1	extractDomainName
E540ES40	String	0	100	<WORD>@<WORD>.<WORD>	emailAddress	-1	-1	strikeoutEmailValidation
F633F623	String	0	96.79	<WORD>	None	-1	-1	expr:CONCATENATE
L2351235	String	0	100	<WORD>	None	-1	-1	expr:CONCATENATE
P500P500	String	0	100	<NUMBER>-<NUMBER>-<NUMBER>	None	-1	-1	splitText
E540ES40	String	0	100	<WORD>@<WORD>.<WORD>	emailAddress	-1	-1	strikeoutEmailValidation
L2361236	Datetime	0	51.2	<NUMBER>/<NUMBER>/<NUMBER>	Date	1.42E+12	1.39E+12	extractFiscalQuarterAndYear
L2361236	Integer	0	0.8	<NUMBER>	None	4	1	expr:CONCATENATE
Q150Q130	String	0	0.8	<WORD>	agreement	-1	-1	Agg
S332S332	String	0	0.8	<WORD>-<WORD>	None	-1	-1	Agg
C530C550	String	0	97.79	<NUMBER>-<NUMBER>-<NUMBER>	None	-1	-1	splitText

FIG. 3

【図４】

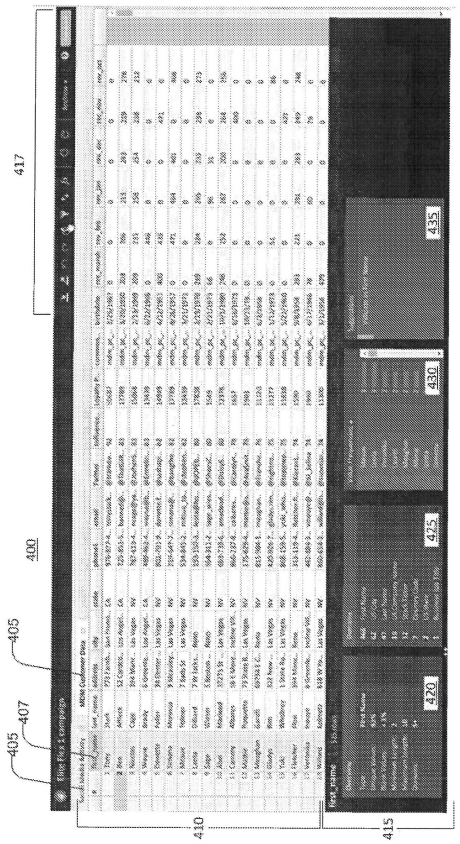


FIG. 4

【図５Ａ】

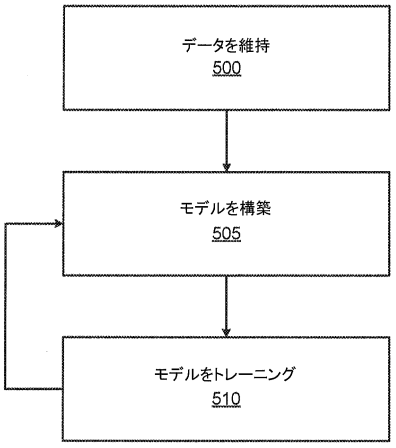


FIG. 5A

【図５Ｂ】

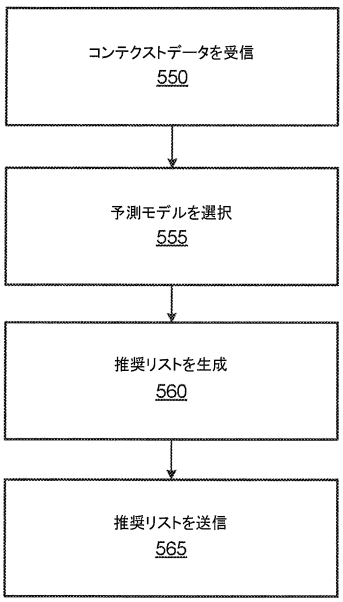


FIG. 5B

【図６】

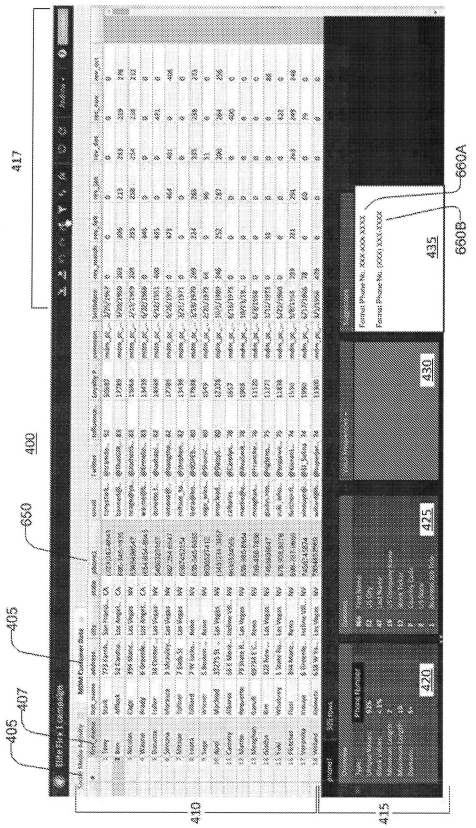


FIG. 6

【図７】

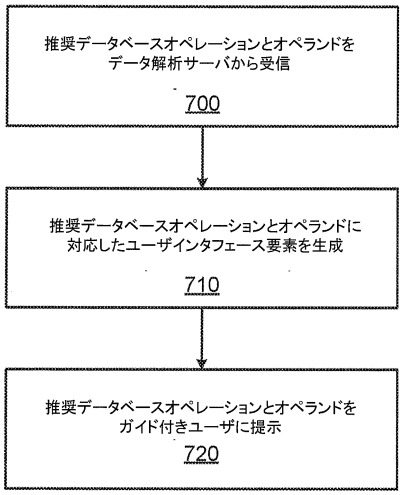


FIG. 7

フロントページの続き

(72)発明者　デイ、　アトレイー
インド国　５６００７７　バンガロール、　コサヌアー、　バイラシ　クロス、　ソリティア　レジデンシー　Ｂ２０１
(72)発明者　カルスカル、　サンジェイ
アメリカ合衆国　９４０６３　カリフォルニア州、　レッドウッド　シティ、　シーポート　ブルーバード　２１００
(72)発明者　ダーンシング、ウダヤクマール
アメリカ合衆国　９５０１４　カリフォルニア州、　クパチーノ、　アパートメント　６４５、バレー　グリーン　ドライブ　２０９９０

審査官　齊藤　貴孝

(56)参考文献　米国特許出願公開第２００７／０２７６８５７（ＵＳ，Ａ１）
特開２０１０－０３３３７７（ＪＰ，Ａ）
中国特許出願公開第１０１８８４０４４（ＣＮ，Ａ）
米国特許出願公開第２０１５／０３２４３４６（ＵＳ，Ａ１）
米国特許出願公開第２０１６／００９２４７５（ＵＳ，Ａ１）
濱崎　雅弘，Ｌｉｎｋｅｄ　Ｏｐｅｎ　ＤａｔａのためのＳＰＡＲＱＬクエリ共有システムの提案，一般社団法人　人工知能学会　第２９回全国大会論文集ＣＤ－ＲＯＭ　［ＣＤ－ＲＯＭ］　２０１５年度　人工知能学会全国大会（第２９回）論文集，日本，社団法人人工知能学会，２０１５年　６月　２日，ｐ．１－４

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　１６／００－１６／９５８