(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2020-161147
(P2020-161147A)
(43)公開日　令和2年10月1日(2020.10.1)

(51)Int.Cl.		F I		テーマコード（参考）
G06F 16/21	(2019.01)	G06F	16/21	5B175

審査請求 有　請求項の数 40　ＯＬ　外国語出願　（全 40 頁）

(21)出願番号　特願2020-88498(P2020-88498)
(22)出願日　令和2年5月20日(2020.5.20)
(62)分割の表示　特願2017-559576(P2017-559576)の分割
原出願日　平成28年6月10日(2016.6.10)
(31)優先権主張番号　62/174,997
(32)優先日　平成27年6月12日(2015.6.12)
(33)優先権主張国・地域又は機関　米国(US)
(31)優先権主張番号　15/175,793
(32)優先日　平成28年6月7日(2016.6.7)
(33)優先権主張国・地域又は機関　米国(US)

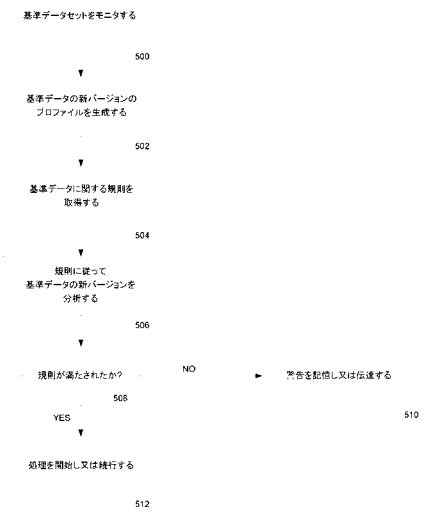
（特許庁注：以下のものは登録商標）

(71)出願人　509123208
アビニシオ　テクノロジー　エルエルシー
アメリカ合衆国　０２４２１　マサチューセッツ州　レキシントン　スプリング　ストリート　２０１
(74)代理人　100079108
弁理士　稲葉　良幸
(74)代理人　100109346
弁理士　大貫　敏史
(74)代理人　100117189
弁理士　江口　昭彦
(74)代理人　100134120
弁理士　内藤　和彦

最終頁に続く

(54)【発明の名称】データ品質分析

(57)【要約】　（修正有）
【課題】データ品質分析の向上を図る方法、非一時的コンピュータ可読媒体及び計算システムを提供する。
【解決手段】方法は、データセットを分析することであって、データセットのフィールドについての基準プロファイルを決定することと、決定した基準プロファイルに基づいて、データセットのフィールドについてのデータ品質規則を決定することと、を含む。データセットのフィールドについてのデータ品質規則は、フィールドについての基準プロファイルと、データセットの一つ以上のデータ記録のフィールドについてのプロファイルとの間の許容偏差、データセットのデータ記録のフィールドのデータ要素についての許容値又はデータセットのデータ記録のフィールドのデータ要素についての禁止値、の一つ以上を示す。
【選択図】図１２



【特許請求の範囲】
【請求項１】
データ処理システムによって生成される出力データセットを示す情報を受信することと、
前記出力データセットに関係するデータ系列情報に基づき、前記出力データセットが依拠する１つ又は複数のアップストリームデータセットを識別することと、
前記出力データセットが依拠する前記識別された１つ又は複数のアップストリームデータセットの１つ又は複数を分析することであって、前記１つ又は複数のアップストリームデータセットのうちの特定のアップストリームデータセットごとに、
（ｉ）前記特定のアップストリームデータセットのプロファイルと前記特定のアップストリームデータセットに関する基準プロファイルとの間の許容偏差を示す第１の規則、及び
（ｉｉ）前記特定のアップストリームデータセット内の１つ又は複数のデータ要素のそれぞれに関する１つ又は複数の許容値又は禁止値を示す第２の規則
のうちの１つ又は複数を適用し、前記１つ又は複数の規則を適用した結果に基づき、前記アップストリームデータセットの１つ又は複数を選択すること、
を含む、分析することと、
前記選択された１つ又は複数のアップストリームデータセットに関連する情報を出力することと、
を含む、方法。
【請求項２】
前記第１の規則及び前記第２の規則の１つ又は複数が自動で生成される、請求項１に記載の方法。
【請求項３】
前記第１の規則が、前記特定のアップストリームデータセットの履歴プロファイルの自動分析に基づいて自動で生成される、請求項２に記載の方法。
【請求項４】
前記基準プロファイルが、前記特定のアップストリームデータセットに関する履歴平均プロファイルに基づく、請求項３に記載の方法。
【請求項５】
前記第２の規則が、前記特定のアップストリームデータセット内の前記１つ又は複数のデータ要素に関する履歴値の自動分析に基づいて自動で生成される、請求項２に記載の方法。
【請求項６】
前記許容値又は禁止値が前記自動分析に基づいて決定される、請求項５に記載の方法。
【請求項７】
前記第１の規則及び前記第２の規則の１つ又は複数がユーザによって指定される、請求項１に記載の方法。
【請求項８】
前記第１の規則及び前記第２の規則の１つ又は複数の指定を、ユーザインタフェースを通して受信することを含む、請求項１に記載の方法。
【請求項９】
データ系列情報が、前記出力データセットが依拠する１つ又は複数のデータセット、前記出力データセットに依拠する１つ又は複数のデータセット、又はその両方を示す、請求項１に記載の方法。
【請求項１０】
前記データセットの部分集合を識別するために前記１つ又は複数のデータセットのそれぞれを分析することが、前記１つ又は複数のデータセットの何れが誤り又は起こり得る誤りを有するかを判定することを含み、
前記方法が、前記部分集合に関して誤り又は起こり得る誤りを有する前記データセット

を選択することを含む、請求項１に記載の方法。
【請求項１１】
前記データセットの部分集合を識別するために前記１つ又は複数のデータセットのそれぞれを分析することが、特定のデータセットであって、前記特定のデータセットの前記プロファイルと前記特定のデータセットに関する前記基準プロファイルとの間の偏差が、前記対応する第１の規則によって示される前記許容偏差を上回る、特定のデータセットを識別することを含み、
前記方法が、前記部分集合のために前記特定のデータセットを選択することを含む、請求項１に記載の方法。
【請求項１２】
前記データセットの部分集合を識別するために前記１つ又は複数のデータセットのそれぞれを分析することが、前記対応する第２の規則によって示される前記許容値又は禁止値を満たさない値を有するデータ要素を有する特定のデータセットを識別することを含み、
前記方法が、前記部分集合のために前記特定のデータセットを選択することを含む、請求項１に記載の方法。
【請求項１３】
前記出力データセット内のデータ要素を識別することを含み、前記出力データセットが依拠する前記１つ又は複数のデータセットを識別することが、前記出力データセット内の前記識別されたデータ要素に影響を及ぼすデータセットを識別することを含む、請求項１に記載の方法。
【請求項１４】
前記出力データセット内のデータ要素を識別することが、誤り又は起こり得る誤りを有するデータ要素を識別することを含む、請求項１３に記載の方法。
【請求項１５】
前記アップストリームデータセットの１つ又は複数のプロファイルを生成することを含む、請求項１に記載の方法。
【請求項１６】
特定のデータセットのプロファイルを生成することが、前記特定のデータセットの新バージョンが受信されるときに前記特定のデータセットの新規プロファイルを生成することを含む、請求項１５に記載の方法。
【請求項１７】
特定のデータセットに関する前記基準プロファイルが、前記特定のデータセットの１つ又は複数の過去のプロファイルから導出される、請求項１に記載の方法。
【請求項１８】
前記データセットの部分集合に関連する情報を出力することが、前記部分集合の前記データセットのそれぞれの識別子を出力することを含む、請求項１に記載の方法。
【請求項１９】
前記データセットの部分集合に関連する情報を出力することが、前記部分集合の前記データセットのそれぞれに関連する誤り又は起こり得る誤りの標識を出力することを含む、請求項１に記載の方法。
【請求項２０】
前記データ処理システムの表現をユーザインタフェース上で表示することを含み、前記データセットの部分集合に関連する情報を出力することが、前記データセットの部分集合の特定のデータセットの表現の近くに前記部分集合の前記特定のデータセットに関連する情報を表示することを含む、請求項１に記載の方法。
【請求項２１】
前記部分集合の前記特定のデータセットに関連する前記表示された情報が、前記特定のデータセットの前記プロファイルと前記特定のデータセットに関する前記基準プロファイルとの間の偏差を示す値を含む、請求項２０に記載の方法。
【請求項２２】

前記部分集合の前記特定のデータセットに関連する前記表示された情報が、前記対応する第２の規則によって示される前記許容値又は禁止値を満たさない前記特定のデータセット内のデータ要素の数を表す値を含む、請求項２０に記載の方法。
【請求項２３】
前記データセットの部分集合に関する情報を示す情報バブル又はポップアップウィンドウを表示することを含む、請求項２０に記載の方法。
【請求項２４】
ユーザが規則を追加するか、規則を修正するか、又は規則を除去することを可能にするためのユーザインタフェースを提供することを含む、請求項１に記載の方法。
【請求項２５】
前記データセットが１つ又は複数のソースデータセット及び１つ又は複数の基準データセットを含み、前記ソースデータセットが、前記データ処理システムによって処理されるデータ要素を含み、前記基準データセットが、前記ソースデータセット内の前記データ要素を処理する際に前記データ処理システムによって参照される基準値を含む、請求項１に記載の方法。
【請求項２６】
前記基準データセットが、前記データ処理システムに関連する企業体に関連するデータを含み、及び前記ソースデータセットが、前記企業体の顧客に関連するデータを含む、請求項２５に記載の方法。
【請求項２７】
前記データ処理システムが変換要素を含み、及び前記方法が、前記出力データセットに影響を及ぼす１つ又は複数の変換要素を前記データ系列情報に基づいて識別することを含む、請求項１に記載の方法。
【請求項２８】
前記変換要素の何れの１つ又は複数が誤り又は起こり得る誤りを有するかを判定することを含む、請求項２７に記載の方法。
【請求項２９】
特定のデータ処理要素が誤り又は起こり得る誤りを有するかどうかを、前記特定の変換要素に関連する実装日に基づいて判定することを含む、請求項２８に記載の方法。
【請求項３０】
データ処理システムによって生成される出力データセットを示す情報を受信することと、
前記出力データセットに関係するデータ系列情報に基づき、前記出力データセットが依拠する１つ又は複数のアップストリームデータセットを識別することと、
前記出力データセットが依拠する前記識別された１つ又は複数のアップストリームデータセットの１つ又は複数を分析することであって、前記１つ又は複数のアップストリームデータセットのうちの特定のアップストリームデータセットごとに、
（ｉ）前記特定のアップストリームデータセットのプロファイルと前記特定のアップストリームデータセットに関する基準プロファイルとの間の許容偏差を示す第１の規則、及び
（ｉｉ）前記特定のアップストリームデータセット内の１つ又は複数のデータ要素のそれぞれに関する１つ又は複数の許容値又は禁止値を示す第２の規則
のうちの１つ又は複数を適用し、前記１つ又は複数の規則を適用した結果に基づき、前記アップストリームデータセットの１つ又は複数を選択すること、
を含む、分析することと、
前記選択された１つ又は複数のアップストリームデータセットに関連する情報を出力することと、
を計算システムに行わせるための命令を記憶する、非一時的コンピュータ可読媒体。
【請求項３１】
メモリに結合されるプロセッサを含む計算システムであって、前記プロセッサ及びメモ

リは、
データ処理システムによって生成される出力データセットを示す情報を受信することと、
前記出力データセットに関係するデータ系列情報に基づき、前記出力データセットが依拠する１つ又は複数のアップストリームデータセットを識別することと、
前記出力データセットが依拠する前記識別された１つ又は複数のアップストリームデータセットの１つ又は複数を分析することであって、前記１つ又は複数のアップストリームデータセットのうちの特定のアップストリームデータセットごとに、
（ｉ）前記特定のアップストリームデータセットのプロファイルと前記特定のアップストリームデータセットに関する基準プロファイルとの間の許容偏差を示す第１の規則、及び
（ｉｉ）前記特定のアップストリームデータセット内の１つ又は複数のデータ要素のそれぞれに関する１つ又は複数の許容値又は禁止値を示す第２の規則
のうちの１つ又は複数を適用し、前記１つ又は複数の規則を適用した結果に基づき、前記アップストリームデータセットの１つ又は複数を選択すること、
を含む、分析することと、
前記選択された１つ又は複数のアップストリームデータセットに関連する情報を出力することと、
を行うように構成される、計算システム。
【請求項３２】
データ処理システムによって生成される出力データセットを示す情報を受信するための手段と、
前記出力データセットに関係するデータ系列情報に基づき、前記出力データセットが依拠する１つ又は複数のアップストリームデータセットを識別するための手段と、
前記出力データセットが依拠する前記識別された１つ又は複数のアップストリームデータセットの１つ又は複数を分析するための手段であって、前記分析することが、前記１つ又は複数のアップストリームデータセットのうちの特定のアップストリームデータセットごとに、
（ｉ）前記特定のアップストリームデータセットのプロファイルと前記特定のアップストリームデータセットに関する基準プロファイルとの間の許容偏差を示す第１の規則、及び
（ｉｉ）前記特定のアップストリームデータセット内の１つ又は複数のデータ要素のそれぞれに関する１つ又は複数の許容値又は禁止値を示す第２の規則
のうちの１つ又は複数を適用し、前記１つ又は複数の規則を適用した結果に基づき、前記アップストリームデータセットの１つ又は複数を選択すること、
を含む、手段と、
前記選択された１つ又は複数のアップストリームデータセットに関連する情報を出力するための手段と、
を含む、計算システム。
【請求項３３】
データ処理システムのダウンストリームデータセットのデータ要素内の誤り又は起こり得る誤りを識別すると、前記ダウンストリームデータセットに関係するデータ系列情報に基づき、前記データ要素に影響を及ぼす１つ又は複数のアップストリームデータセットを自動で識別することと、
前記識別されたアップストリームデータセットのそれぞれの現在のプロファイル及び基準プロファイルを分析することを含む、前記アップストリームデータセットの何れが誤り又は起こり得る誤りを有するかを判定することと、
誤りを有するか又は誤りを有する可能性が高いと判定される前記アップストリームデータセットのそれぞれに関連する情報を出力することと、
を含む、方法。

【発明の詳細な説明】
【技術分野】
【０００１】
背景
本明細書は、データ品質分析に関する。データセットのデータ品質は、データセット内のデータ記録が誤りを有するかどうかの指標である。多くの場合、データセットの処理中に誤りが生じる場合、そのデータセットのデータ品質は低い。
【発明の概要】
【課題を解決するための手段】
【０００２】
要約
一般的な態様では、方法は、データ処理システムによって生成される出力データセットを示す情報を受信することと、出力データセットに関係するデータ系列情報に基づき、出力データセットが依拠する１つ又は複数のアップストリームデータセットを識別することと、出力データセットが依拠する識別された１つ又は複数のアップストリームデータセットの１つ又は複数を分析することとを含む。分析することは、１つ又は複数のアップストリームデータセットのうちの特定のアップストリームデータセットごとに、（ｉ）特定のアップストリームデータセットのプロファイルと特定のアップストリームデータセットに関する基準プロファイルとの間の許容偏差を示す第１の規則、及び（ｉｉ）特定のアップストリームデータセット内の１つ又は複数のデータ要素のそれぞれに関する１つ又は複数の許容値又は禁止値を示す第２の規則のうちの１つ又は複数を適用し、且つ１つ又は複数の規則を適用した結果に基づき、アップストリームデータセットの１つ又は複数を選択することを含む。この方法は、選択された１つ又は複数のアップストリームデータセットに関連する情報を出力することを含む。
【０００３】
実施形態は、以下の特徴の１つ又は複数を含み得る。
【０００４】
第１の規則及び第２の規則の１つ又は複数が自動で生成される。第１の規則は、特定のアップストリームデータセットの履歴プロファイルの自動分析に基づいて自動で生成される。基準プロファイルは、特定のアップストリームデータセットに関する履歴平均プロファイルに基づく。第２の規則は、特定のアップストリームデータセット内の１つ又は複数のデータ要素に関する履歴値の自動分析に基づいて自動で生成される。許容値又は禁止値は自動分析に基づいて決定される。
【０００５】
第１の規則及び第２の規則の１つ又は複数がユーザによって指定される。
【０００６】
この方法は、第１の規則及び第２の規則の１つ又は複数の指定を、ユーザインタフェースを通して受信することを含む。
【０００７】
データ系列情報は、出力データセットが依拠する１つ又は複数のデータセット、出力データセットに依拠する１つ又は複数のデータセット、又はその両方を示す。
【０００８】
データセットの部分集合を識別するために１つ又は複数のデータセットのそれぞれを分析することは、１つ又は複数のデータセットの何れが誤り又は起こり得る誤りを有するかを判定することを含み、この方法は、部分集合に関して誤り又は起こり得る誤りを有するデータセットを選択することを含む。
【０００９】
データセットの部分集合を識別するために１つ又は複数のデータセットのそれぞれを分析することは、特定のデータセットであって、特定のデータセットのプロファイルと特定のデータセットに関する基準プロファイルとの間の偏差が、対応する第１の規則によって

示される許容偏差を上回る、特定のデータセットを識別することを含み、この方法は、部分集合のために特定のデータセットを選択することを含む。

【００１０】

データセットの部分集合を識別するために１つ又は複数のデータセットのそれぞれを分析することは、対応する第２の規則によって示される許容値又は禁止値を満たさない値を有するデータ要素を有する特定のデータセットを識別することを含み、この方法は、部分集合のために特定のデータセットを選択することを含む。

【００１１】

この方法は、出力データセット内のデータ要素を識別することを含み、出力データセットが依拠する１つ又は複数のデータセットを識別することは、出力データセット内の識別されたデータ要素に影響を及ぼすデータセットを識別することを含む。出力データセット内のデータ要素を識別することは、誤り又は起こり得る誤りを有するデータ要素を識別することを含む。

【００１２】

この方法は、アップストリームデータセットの１つ又は複数のプロファイルを生成することを含む。特定のデータセットのプロファイルを生成することは、特定のデータセットの新バージョンが受信されるときに特定のデータセットの新規プロファイルを生成することを含む。

【００１３】

特定のデータセットに関する基準プロファイルは、特定のデータセットの１つ又は複数の過去のプロファイルから導出される。

【００１４】

データセットの部分集合に関連する情報を出力することは、部分集合のデータセットのそれぞれの識別子を出力することを含む。

【００１５】

データセットの部分集合に関連する情報を出力することは、部分集合のデータセットのそれぞれに関連する誤り又は起こり得る誤りの標識を出力することを含む。

【００１６】

この方法は、データ処理システムの表現をユーザインタフェース上で表示することを含み、データセットの部分集合に関連する情報を出力することは、データセットの部分集合の特定のデータセットの表現の近くに部分集合の特定のデータセットに関連する情報を表示することを含む。部分集合の特定のデータセットに関連する表示された情報は、特定のデータセットのプロファイルと特定のデータセットに関する基準プロファイルとの間の偏差を示す値を含む。部分集合の特定のデータセットに関連する表示された情報は、対応する第２の規則によって示される許容値又は禁止値を満たさない特定のデータセット内のデータ要素の数を表す値を含む。この方法は、データセットの部分集合に関する情報を示す情報バブル又はポップアップウィンドウを表示することを含む。

【００１７】

この方法は、ユーザが規則を追加するか、規則を修正するか、又は規則を除去することを可能にするためのユーザインタフェースを提供することを含む。

【００１８】

データセットは１つ又は複数のソースデータセット及び１つ又は複数の基準データセットを含み、ソースデータセットは、データ処理システムによって処理されるデータ要素を含み、基準データセットは、ソースデータセット内のデータ要素を処理する際にデータ処理システムによって参照される基準値を含む。基準データセットは、データ処理システムに関連する企業体に関連するデータを含み、及びソースデータセットは、企業体の顧客に関連するデータを含む。

【００１９】

データ処理システムは変換要素を含み、及びこの方法は、出力データセットに影響を及ぼす１つ又は複数の変換要素をデータ系列情報に基づいて識別することを含む。この方法

は、変換要素の何れの１つ又は複数が誤り又は起こり得る誤りを有するかを判定することを含む。この方法は、特定のデータ処理要素が誤り又は起こり得る誤りを有するかどうかを、特定の変換要素に関連する実装日に基づいて判定することを含む。

【００２０】

一般的な態様では、非一時的コンピュータ可読媒体は、データ処理システムによって生成される出力データセットを示す情報を受信することと、出力データセットに関係するデータ系列情報に基づき、出力データセットが依拠する１つ又は複数のアップストリームデータセットを識別することと、出力データセットが依拠する識別された１つ又は複数のアップストリームデータセットの１つ又は複数を分析することとを計算システムに行わせるための命令を記憶する。分析することは、１つ又は複数のアップストリームデータセットのうちの特定のアップストリームデータセットごとに、（ｉ）特定のアップストリームデータセットのプロファイルと特定のアップストリームデータセットに関する基準プロファイルとの間の許容偏差を示す第１の規則、及び（ｉｉ）特定のアップストリームデータセット内の１つ又は複数のデータ要素のそれぞれに関する１つ又は複数の許容値又は禁止値を示す第２の規則のうちの１つ又は複数を適用し、且つ１つ又は複数の規則を適用した結果に基づき、アップストリームデータセットの１つ又は複数を選択することを含む。命令は、選択された１つ又は複数のアップストリームデータセットに関連する情報を計算システムに出力させる。

【００２１】

一般的な態様では、計算システムは、メモリに結合されるプロセッサを含む。プロセッサ及びメモリは、データ処理システムによって生成される出力データセットを示す情報を受信することと、出力データセットに関係するデータ系列情報に基づき、出力データセットが依拠する１つ又は複数のアップストリームデータセットを識別することと、出力データセットが依拠する識別された１つ又は複数のアップストリームデータセットの１つ又は複数を分析することとを行うように構成される。分析することは、１つ又は複数のアップストリームデータセットのうちの特定のアップストリームデータセットごとに、（ｉ）特定のアップストリームデータセットのプロファイルと特定のアップストリームデータセットに関する基準プロファイルとの間の許容偏差を示す第１の規則、及び（ｉｉ）特定のアップストリームデータセット内の１つ又は複数のデータ要素のそれぞれに関する１つ又は複数の許容値又は禁止値を示す第２の規則のうちの１つ又は複数を適用し、且つ１つ又は複数の規則を適用した結果に基づき、アップストリームデータセットの１つ又は複数を選択することを含む。プロセッサ及びメモリは、選択された１つ又は複数のアップストリームデータセットに関連する情報を出力するように構成される。

【００２２】

一般的な態様では、計算システムは、データ処理システムによって生成される出力データセットを示す情報を受信するための手段と、出力データセットに関係するデータ系列情報に基づき、出力データセットが依拠する１つ又は複数のアップストリームデータセットを識別するための手段と、出力データセットが依拠する識別された１つ又は複数のアップストリームデータセットの１つ又は複数を分析するための手段とを含む。分析することは、１つ又は複数のアップストリームデータセットのうちの特定のアップストリームデータセットごとに、（ｉ）特定のアップストリームデータセットのプロファイルと特定のアップストリームデータセットに関する基準プロファイルとの間の許容偏差を示す第１の規則、及び（ｉｉ）特定のアップストリームデータセット内の１つ又は複数のデータ要素のそれぞれに関する１つ又は複数の許容値又は禁止値を示す第２の規則のうちの１つ又は複数を適用し、且つ１つ又は複数の規則を適用した結果に基づき、アップストリームデータセットの１つ又は複数を選択することを含む。この計算システムは、選択された１つ又は複数のアップストリームデータセットに関連する情報を出力するための手段を含む。

【００２３】

一般的な態様では、方法は、データ処理システムのダウンストリームデータセットのデータ要素内の誤り又は起こり得る誤りを識別すると、ダウンストリームデータセットに関

係するデータ系列情報に基づき、データ要素に影響を及ぼす１つ又は複数のアップストリームデータセットを自動で識別することと、識別されたアップストリームデータセットのそれぞれの現在のプロファイル及び基準プロファイルを分析することを含む、アップストリームデータセットの何れが誤り又は起こり得る誤りを有するかを判定することと、誤りを有するか又は誤りを有する可能性が高いと判定されるアップストリームデータセットのそれぞれに関連する情報を出力することとを含む。

【００２４】

態様は以下の利点の１つ又は複数を含み得る。

【００２５】

本明細書に記載の手法は、データ分析官又はアプリケーション開発者等のユーザがデータ品質問題の根本的原因を迅速に識別することを補助し得る。例えば、データ処理システム内の基準データは頻繁に更新されるが、必ずしも導入前に完全に検査されない可能性がある。基準データ内の誤りは、基準データを使用して処理されるダウンストリームデータ内のデータ品質問題を引き起こし得る。ダウンストリームデータセット内のデータ品質問題の根本的原因を分析することは、ダウンストリームデータセットのデータ品質に影響を及ぼしている可能性があるデータ品質問題を有する基準データ又は他のアップストリームデータを識別するのを補助し得る。潜在的なデータ品質問題をユーザに通知することは、ユーザがデータ処理をプロアクティブに管理することを補助し得る。

【００２６】

本発明の他の特徴及び利点が以下の説明及び特許請求の範囲から明らかになる。

【図面の簡単な説明】

【００２７】

【図１】データ系列図である。
【図２】データ系列図である。
【図３Ａ】データ系列図である。
【図３Ｂ】データ系列図である。
【図４】ユーザインタフェースの図である。
【図５】システム図である。
【図６】ユーザインタフェースの図である。
【図７】データ処理システムの図である。
【図８Ａ】データ処理システムの図である。
【図８Ｂ】データ処理システムの図である。
【図８Ｃ】記録の一例である。
【図９Ａ】データ処理システムの図である。
【図９Ｂ】データ処理システムの図である。
【図１０Ａ】データ処理システムの図である。
【図１０Ｂ】記録の一例である。
【図１１】フローチャートである。
【図１２】フローチャートである。
【図１３】フローチャートである。
【図１４】フローチャートである。
【図１５】フローチャートである。
【図１６】システム図である。

【発明を実施するための形態】

【００２８】

説明

データ系列の分析に基づいてデータ品質問題の根本的原因を識別するための手法をここで説明する。データ品質問題がダウンストリームデータセット内で識別される場合、ダウンストリームデータセットの導出元であるアップストリームデータセット及びアップストリーム変換要素（アップストリームデータ系列要素と呼ばれる場合もある）が識別される

。ダウンストリームデータセット内のデータ品質問題に寄与したデータ品質問題をそれ自体が有し得るアップストリームデータ系列要素の１つ又は複数を識別するために、各アップストリームデータ系列要素の品質が評価される。一部の例では、データセットがデータ品質問題を有するかどうかを判定するために、各アップストリームデータセットを特徴付けるプロファイルがそのデータセットに関する履歴平均プロファイル等の基準プロファイルと比較される。一部の例では、データセットがデータ品質問題を有するかどうかを判定するために、アップストリームデータセットのフィールド内の値がそのフィールドに関する１つ又は複数の許容値又は禁止値と比較される。

【００２９】

データ系列とは、データ処理システムによって処理されるデータ記録のライフサイクルを記述する情報である。所与のデータセットのためのデータ系列情報は、所与のデータセットが依拠する１つ又は複数のアップストリームデータセットの識別子、所与のデータセットに依拠する１つ又は複数のダウンストリームデータセット、及びデータを処理して所与のデータセットを生成する１つ又は複数の変換を含む。アップストリームデータセットに依拠するダウンストリームデータセットとは、データ処理システムによるアップストリームデータセットの処理がダウンストリームデータセットの生成を直接又は間接的にもたらすことを意味する。生成されるダウンストリームデータセットは、データ処理システムから出力されるデータセット（出力データセットと呼ばれる場合もある）とすることができ、又はデータ処理システムによって更に処理されるデータセット（中間データセットと呼ばれる場合もある）とすることができる。アップストリームデータセットは、データ処理システム内に入力されるデータセット（入力データセット又は基準データセットと呼ばれる場合もある）、又はデータ処理システムによって既に処理されているデータセット（中間データセットと呼ばれる場合もある）とすることができる。変換とは、データシンクに与えられるダウンストリームデータセットをもたらすためにアップストリームデータセットに適用されるデータ処理操作である。データ系列図は、データ処理システム内のデータ系列要素のグラフィカル表現である。

【００３０】

図１は、データ処理システムによって生成される出力データ１１０に関するデータ系列図１００の一例である。図１の例では、データ処理システムがソースデータ１０２、１０４の２つのセットを受信する。ソースデータは、例えば計算システム内のデータを記憶するための単層ファイル等のファイル、リレーショナルデータベース若しくはオブジェクトデータベース等のデータベース、待ち行列若しくは別のリポジトリ内に記憶されるデータ記録、又はそれらから受信されるデータ記録とすることができる。例えば、ソースデータ１０２は、ファイル「US_feed.dat」内に記憶された米国内でのクレジットカード取引のデータ記録とすることができる。各データ記録は、記録構造内で定められる属性又はデータベーステーブル内のカラム等、１つ又は複数のフィールドのそれぞれのための値を含むことができる。ソースデータ１０２、１０４はバッチ単位で受信され処理され得る（例えば、毎時、毎日、毎週、毎月、毎四半期、毎年、又は別の間隔で処理されるファイル又はデータベースからのデータ）。ソースデータ１０２、１０４はストリームとして受信され、継続的に処理されてもよく、例えば待ち行列によってバッファされ、データが入手可能であり且つシステム資源が許すときに処理され得る。

【００３１】

ソースデータ１０２は変換要素１０６によって処理され、変換要素１０６は、例えばソースデータ１０２を何らかの方法で変えるためにソースデータ１０２に作用する。変換要素は、仮想マシン内で実行されるjavaプログラム、実行ファイル、データフローグラフ、別の種類の実行可能プログラム等、データを操作可能な実行可能プログラムとすることができる。例えば、変換要素１０６は「TransformA.exe」と名付けられた実行ファイルであり得る。ある具体例では、変換要素１０６が、正しくないフォーマットを有するデータ記録等、ソースデータ１０２から不所望のデータ記録をフィルタで除去するフィルタコンポーネントであり得る。変換要素１０６は、基準データ１２０を考慮してソースデータ１０

２を処理して中間データ１１２をもたらす。基準データは、変換要素がデータを処理することを可能にするために変換要素によって使用されるデータである。例えば、マッピング操作を可能にする基準データは、処理されているデータ内の１つ又は複数のフィールド内の値に対応する値を有する１つ又は複数のフィールドを含む。中間データ１１２は、計算システム内のデータを記憶するためのファイル、データベース、待ち行列、又は別のリポジトリ内に記憶され得る。

【００３２】

変換要素１０８は、基準データ１２２を考慮してソースデータ１０４のセットを処理して中間データ１１４をもたらす。中間データ１１４は、計算システム内のデータを記憶するためのファイル、データベース、待ち行列、又は別のリポジトリ内に記憶され得る。

【００３３】

中間データ１１２、１１４は、基準データ１１８を使用する変換要素１１６によって一緒に処理される。一例では、変換要素１１６がマッピング操作であり、基準データ１１８が州の値及び対応する地域の値を示すデータ記録を含む。中間データ１１２、１１４が変換要素１１６によって処理されるとき、中間データ１１２、１１４内の各データ記録内の州フィールド内の値が基準データ１１８内で示される対応する地域にマップされる。一例では、基準データ１１８が、法人企業体、対応する部門識別子、経営者名、及び位置を示すビジネスデータを含む。中間データ１１２、１１４が変換要素１１６によって処理されるとき、基準データセットによって可能にされるマッピングに基づいて各データ記録が法人企業体に割り振られる。基準データ１１８は複数のデータセットを処理するために使用することができ、処理によって変更されない。基準データ１１８はユーザによって周期的に又は必要に応じて更新され得る。

【００３４】

変換要素１１６は、計算システム内のデータを記憶するためのファイル、データベース、待ち行列、又は別のリポジトリ内に記憶される出力データ１１０を出力する。出力データ１１０は、例えば同じデータ処理システム内の若しくは異なるデータ処理システム内の他の変換要素によって更に処理されてもよく、又は将来分析するために記憶され得る。

【００３５】

図１の例では、単一のデータ処理システム内のデータ系列要素に関して出力データ１１０のデータ系列が図示されている。一部の例では、複数のデータ処理システムによってデータセットのデータ系列を追跡することができる。例えば、出力データＸをもたらすためにソースデータが第１のデータ処理システムによって最初に処理され得る。第２のデータ処理システムが、第１のデータ処理システムからの出力データＸを読み取り、出力データＸを処理して出力データＹを生成する。出力データＹは第３のデータ処理システムによって処理され、出力データＺが生成される。出力データＺのデータ系列は、最初のソースデータ、３つのデータ処理システムのそれぞれに含まれる変換、及び３つのデータ処理システムの何れかによる処理中に使用される任意の基準データを含む。

【００３６】

一部の例では、目標要素２０６Ａのための終端間データ系列図２００Ａの例で図示されているような、より複雑なデータ処理システムによって出力データが生成され得る。データ系列図２００Ａでは、データ要素２０２Ａと変換要素２０４Ａとの間のつながりが図示されている。データ要素２０２Ａはデータセット、データセット内のテーブル、テーブル内のカラム、ファイル内のフィールド、又は他のデータを表し得る。変換要素の一例は、データ要素の単一出力がどのように作り出されるかを記述する実行ファイルの要素である。図２のデータ処理システム内で目標要素２０６Ａ内の（又は別のデータ要素２０２Ａ内の）潜在的なデータ品質問題の根本的原因を追跡することができる。図２の更なる説明は、参照によりその全内容を本明細書に援用する米国特許出願公開第２０１０／０１３８４３１号に見出すことができる。

【００３７】

図１又は図２のデータ系列図等のデータ系列図内で示されている情報は、何れのアップ

ストリームデータソース、データシンク、又は変換がダウンストリームデータに影響を及ぼすかを示す。例えば、図１のデータ系列図１００は、出力データ１１０がソースデータ１０２、１０４、基準データ１１８、１２０、１２２、及び変換要素１０６、１０８、１１６の影響を受けていることを明らかにする。

【００３８】

ダウンストリームデータセット（出力データ１１０等）の系列を理解することは、ダウンストリームデータ内で生じ得るデータ品質問題の根本的原因を識別する際に有用であり得る。データ品質問題の根本的原因とは、ダウンストリームデータ内のデータ品質問題の少なくとも部分的な原因であるアップストリームのシステム、操作、又はデータセットの識別を意味する。出力データ１１０等におけるダウンストリームデータセット内のデータ品質問題の原因は、低品質のソースデータ、低品質の基準データ、出力データ１１０のセットのアップストリーム系列内の変換要素内の誤り、又はそれらのうちの何れか２つ以上の組合せであり得る。データ系列要素の品質又は状態を追跡することは、低品質の出力データのあり得る根本的原因を評価するために使用できる情報をもたらし得る。

【００３９】

データセットのデータ品質とは、そのデータセットが予期される特性を有するかどうかを概して意味する。低いデータ品質は、予期された通りに振る舞わない、例えば統計的標準を外れる、標準的な照会に応答してルックアップの失敗又は別の種類の挙動を返すデータセット内に現れ得る。データセットの品質は、以下で説明するように、データセット内のデータ記録の一部若しくは全てのプロファイルに基づいて、特定のデータ記録の１つ又は複数のフィールドのそれぞれの中の値に基づいて、又はその両方に基づいて特徴付けることができる。

【００４０】

ダウンストリームデータセット（例えば、出力データ１１０）内の低いデータ品質は、出力データのアップストリームデータ系列内の様々な要因の何れかにさかのぼることができる。低品質の出力データの１つのあり得る原因は、低品質のソースデータ、低品質の基準データ、又はその両方であり得る。例えば、あるソースデータセットは、伝送中に破損し若しくは中断されている場合があり、間違ったデータセットである可能性があり、欠落データを有することがあり、又は別の問題を有し得る。基準データセットは、基準データセットに対する最近の更新で誤りにさらされている場合があり、破損していることがあり、間違ったデータセットである可能性があり、又は別の問題を有し得る。低品質の出力データの別のあり得る原因は、出力データのアップストリームデータ系列内の変換要素の問題であり得る。例えば、変換要素を実装するソフトウェアが新バージョンに最近更新された場合、例えば更新されたソフトウェアが誤りを有し又は破損している場合、変換要素はもはや所望の処理を行わない場合がある。出力データセット内で生じ得る潜在的なデータ品質問題を先制して識別すること、出力データセット内で生じたデータ品質問題の根本的原因を後に追跡すること、又はその両方を容易にするために、出力データ１１０のセットのデータ系列内のソースデータ、基準データ、及び変換要素をモニタすることができる。

【００４１】

ソースデータ及び基準データをモニタリングし分析することは、低品質の出力データの１つ又は複数のあり得る原因をユーザが診断するのを補助し得る。例えば、低品質の出力データセットが生成される場合、所与のソースデータセット又は基準データセット自体が低品質かどうか、従って低品質の出力データのあり得る一因かどうかを、低品質の出力データセットのデータ系列内のソースデータ又は基準データを分析することが示し得る。ソースデータ及び基準データをモニタリングすることは、処理された場合にダウンストリーム出力データ内でデータ品質問題を引き起こし得る、低品質のソースデータ又は基準データを先制して識別することもできる。

【００４２】

図３Ａ及び図３Ｂは、図１に示したデータ系列を有する出力データ１１０のセット内の未知の又は潜在的なデータ品質問題の根本的原因を追跡する手法を示す。図３Ａを参照す

ると、入力データ（例えば、図１のソースデータ１０２、１０４）を処理する前に、基準データ１１８、１２０、１２２の品質が品質要素１５４、１５６、１５８によってそれぞれ特徴付けられる。一部の例では、基準データセットが更新されるとき、予定時刻に（例えば、周期的に又は基準データの更新が予定されているとき）、各入力データセットを処理する前に、又は他の時点において基準データの品質を特徴付けることができる。

【００４３】

データセットの品質を特徴付けるために、品質要素がデータセット内のフィールドのプロファイル（統計調査と呼ばれる場合もある）を計算する。データ記録セットのプロファイルは、データ記録内のデータ値の例えばフィールドごとの要約である。プロファイルは、セット内のデータ記録の少なくとも一部のそれぞれの中の１つ又は複数のフィールドのそれぞれの中のデータ値を特徴付ける統計、値のヒストグラム、最大値、最小値、平均（例えば、中間又は中央）値、平均値からの標準偏差、個別値の数、（例えば、データセットごとの重要なデータ要素に関する）１つ又は複数のフィールド内の最も高頻度の値及び最も低頻度の値の標本、又は他の統計を含む。一部の例では、プロファイルが、データ記録内の１つ又は複数のフィールドのそれぞれの中のデータ値を特徴付ける処理された情報を含み得る。例えば、プロファイルは、フィールド内の値の分類（例えば、収入データフィールド内のデータの高い、中位、又は低いカテゴリへの分類）、個々のデータ記録内のデータフィールド間の関係の指示（例えば、州のデータフィールドとＺＩＰのデータフィールドとが無関係ではないという指示）、データ記録間の関係（例えば、顧客＿識別子フィールド内で共通値を有するデータ記録は関係しているという指示）、又はデータ記録セット内のデータを特徴付ける他の情報を含み得る。

【００４４】

次いで、品質要素が１つ又は複数の規則を適用して、データセット内の任意の実際の又は潜在的なデータ品質問題を識別する。以下で更に論じるように、規則はユーザによって指定されてもよく、プロファイルの許容可能な特徴又は禁止された特徴を示し得る。ある具体例では、基準データセットが米国の州の略記を列挙するフィールドを含む場合、規則の一例は、そのフィールド内の個別値の数が５０を上回る場合、データ品質問題を識別すべきであると示すことができる。一部の例では、規則がデータセットの履歴プロファイル、例えば履歴平均値に基づき得る。データ品質問題がデータセット内で識別されない場合、規則を更新するために、例えば履歴平均値を更新するためにデータセットのプロファイルを使用することができる。実際の又は潜在的なデータ品質問題を有するものとして基準データセットが識別される場合、データ品質問題が対処されるまで処理を休止することができる。

【００４５】

図３Ｂを参照すると、ソースデータ１０２、１０４の品質は品質要素１５０、１５２によってそれぞれ特徴付けられる。品質要素１５０、１５２は、データ処理システム内にデータが受信されるとき、それぞれのソースデータの予定された処理の前に、又は他の時点においてソースデータ１０２、１０４のそれぞれのデータ品質を特徴付けることができる。既知の又は潜在的なデータ品質問題を有するものとしてソースデータセットが識別される場合、例えばユーザに警告するために又は将来参照するためにデータ記憶域内に記憶するために、データ品質問題に関する情報を出力することができる。例えば、各品質要素１５０、１５２が対応するデータセットからデータを読み取ると、品質要素１５０、１５２はデータセットのプロファイルを計算する。

【００４６】

ある具体例では、ソースデータ１０２のプロファイルを計算するために、品質要素１５０はソースデータ１０２内の取引＿量フィールド内の値の全ての和を計算することができる。ソースデータ１０２のための規則が、取引＿量フィールド内の値の全ての和を過去３０ランにわたるその和の平均及び標準偏差と比較することができ、ソースデータ１０２の取引＿量フィールド内の値の全ての和が和の平均値からの１標準偏差の範囲外である場合、データ品質問題を識別すべきであると示すことができる。

【００４７】
一部の例では、データセットの品質を特徴付けるために使用される規則が、データセット内のデータ記録のプロファイルの許容可能な特徴又は禁止された特徴を示し得る。プロファイルの特徴は値又は値域とすることができる。プロファイルの許容可能な特徴を示す規則は、プロファイルが許容可能な特徴を含む場合に満たされる。フィールドに関する許容可能な特徴の一例は、そのフィールドの許容可能な最大値及び最小値とすることができ、そのフィールドの平均値が許容可能な最大値と最小値との間に含まれる場合に規則が満たされる。プロファイルの禁止された特徴を示す規則は、禁止された特徴をプロファイルが含まない限り満たされる。フィールドに関する禁止された特徴の一例は、そのフィールドについて禁止されている値の一覧とすることができ、そのフィールドが禁止値の何れかを含む場合、その規則は満たされない。

【００４８】
プロファイルの特徴を示す規則は、特定のデータセットのフィールドのプロファイルとデータセットのフィールドに関する基準プロファイルとの間の許容偏差を示し得る。対応する規則によって示される許容偏差を上回るデータセットのプロファイルとデータセットに関する基準プロファイルとの間の偏差は、そのデータセット内のデータ品質問題の指示、従ってそのデータセットがダウンストリームデータセット内の既存の又は潜在的なデータ品質問題のあり得る根本的原因であるという指示であり得る。一部の例では、最大許容値及び最小許容値等の値域として許容偏差を指定することができる。一部の例では、平均値（例えば、過去のデータセット内の値の中間又は中央）とすることができる単一値からの標準偏差として許容偏差を指定することができる。

【００４９】
一部の例では、データセットの品質を特徴付けるために使用される規則は、フィールド内の値の妥当性等に基づき、データ記録の１つ又は複数のフィールドのそれぞれの中の値の許容特性又は禁止特性を示すことができる。フィールドに関する許容特性を示す規則は、フィールド内の値がその許容特性に適合する場合に満たされる。フィールドに関する禁止特性を示す規則は、フィールド内の値が禁止特性に適合しない限り満たされる。規則を満たす値は有効値と呼ばれる場合もあり、規則を満たさない値は無効値と呼ばれる場合もある。規則による許容特性又は禁止特性としてフィールド内の値の様々な特性を示すことができる。規則の一例は、許容値域若しくは禁止値域、最大許容値、最小許容値、又は許容若しくは禁止されている１つ又は複数の特定の値の一覧等、フィールドのコンテンツの許容特性又は禁止特性を示すことができる。例えば、１９００未満の値又は２０１６を上回る値を有する誕生＿年フィールドは無効と見なされ得る。規則の一例は、フィールドのデータタイプの許容特性又は禁止特性を示すことができる。規則の一例は、特定のフィールド内に値がないこと（又はヌルがあること）が許容されているか、又は禁止されているかを示すことができる。例えば、文字列値（例えば、「Smith」）を含むラスト＿ネームフィールドは有効と見なされ得る一方、空白であり又は数値を含むラスト＿ネームフィールドは無効と見なされ得る。規則の一例は、同じデータ記録内の２つ以上のフィールド間の許容又は禁止された関係を示し得る。例えば、ある規則は、州フィールドのあり得る各値に対応するＺＩＰフィールドの値の一覧を指定することができ、その一覧によってサポートされないＺＩＰフィールドの値と、州フィールドの値とのいかなる組合せも無効であると指定することができる。

【００５０】
一部の例では、履歴データを自動で分析することに基づいて規則を生成することができる。この種の規則を自動生成規則と呼ぶ。自動生成規則は、データセット内のデータ記録のプロファイルの許容可能な特徴又は禁止された特徴を示すことができる。例えば、プロファイルの自動生成規則は、特定のデータセットのフィールドのプロファイルとデータセットのフィールドの自動的に決定された履歴基準プロファイルとの間の許容偏差を示し得る。データセットの履歴基準プロファイルは履歴データに基づくことができ、例えば、履歴基準プロファイルは前日の同じデータセットのプロファイル、過去の複数日の（例えば

、先週又は先月にわたる）同じデータセットの平均プロファイル、同じデータセットの存続期間の平均プロファイルとすることができる。より広義には、基準プロファイルは、様々な種類の統計的分析を活用するための多岐にわたる基準情報を保持することができる。例えば、基準プロファイルは、値分布の標準偏差又は他の指示に関する情報を含み得る。以下の例では、及び本願の一般的な概念を限定することなく、基準プロファイルが過去のデータセットの数値的平均及び場合により標準偏差も含むと仮定する。

【００５１】

自動生成規則は、データ記録のフィールド内の値の自動的に決定された許容特性又は禁止特性を示すことができる。一例では、あるフィールドに関する自動生成規則が、そのフィールドの履歴的な最大値又は最小値の分析に基づくフィールドの許容可能な最大値又は最小値を示し得る。一例では、あるフィールドに関する自動生成規則が、そのフィールドについて過去に生じた値の分析に基づくフィールドの許容値一覧を示し得る。一部の例では、データセットの全フィールドについて自動生成規則が指定される。一部の例では、フィールドの部分集合について規則が指定される。規則が指定されるフィールドは、例えばデータ記録の分析に基づいて自動で識別され得る。例えば、自動生成規則を生成可能なフィールドとして、少数の個別値を概して有するデータ記録セット内の任意のフィールド（低濃度フィールドと呼ばれる場合もある）を識別することができる。

【００５２】

一部の例では、自動生成規則を生成するために機械学習技法が使用される。例えば、規則を生成する前に履歴平均又は期待値を識別するために、学習期間にわたってデータを分析することができる。学習期間は、指定の期間とすることができ、又は平均値若しくは期待値が安定した値に収束するまでの時間とすることができる。

【００５３】

一部の例では、規則がユーザによって指定され得る。この種の規則をユーザ指定規則と呼ぶ。ユーザ指定規則は、特定のデータセットのフィールドのプロファイルの許容特性又は禁止特性、データセット内のデータ記録の１つ又は複数のフィールドのそれぞれの中の値の許容特性又は禁止特性、又はその両方を指定することができる。ユーザは、例えばシステムによって処理されるデータ記録の予期される特性についての自らの理解に基づいて規則を指定することができる。一部の例では、ユーザによって修正され得る省略時値をユーザ指定規則に割り当てることができる。

【００５４】

ある具体例では、ソースデータが米国内で生じる取引に関するクレジットカード取引記録である。ソースデータは、１時間のインクリメント単位で処理されるストリーミングデータである。ソースデータについての、及びクレジットカード取引記録を処理するときに実行すべき操作についての自らの知識に基づき、ユーザは、プロファイルすべき重要なデータ要素として取引識別子フィールド、カード識別子フィールド、州フィールド、日付フィールド、及び金額フィールドを識別することができる。

【００５５】

ソースデータがクレジットカード取引記録である具体例では、ユーザは、州フィールドについて５０個の許容値のみがあることを知っている場合がある。ユーザは、基準に対するソースデータセットのプロファイルの標準偏差に関係なく、ソースデータセットのプロファイルが州フィールド内で５０を上回る値を識別する場合、警告フラグの使用を引き起こす規則を作成することができる。ユーザは、処理と同日に完了した取引に関するクレジットカード取引記録のみがソースデータセット内にあることも知っている場合がある。ユーザは、任意のソースデータ記録が処理日と一致しない日付を有する場合、警告メッセージの送信を引き起こす規則を作成することができる。

【００５６】

図４を参照すると、一部の例では、ユーザが１つ又は複数の規則をユーザインタフェース４００によって指定することができる。ユーザインタフェース４００の一例は、複数の行４０２及び複数の列４０４を含む。各行４０２は、データセット内のデータ記録のフィ

ールド４０６に関連し、各列４０４は規則４０８に関連する。ユーザインタフェース４００により、ユーザは１つ又は複数のフィールド４０６に関する規則を指定することができ、又はフィールドに関する事前にデータ投入された省略時規則を承認することができる。ユーザインタフェース４００についての更なる説明は、参照によりその全内容を本明細書に援用する、２０１２年１０月１７日に出願された米国特許出願第１３／６５３，９９５号に見出すことができる。ユーザインタフェース４００の他の実装形態もあり得る。

【００５７】

一部の例では、基準データセットの新バージョン内又はソースデータセット内等、起こり得るデータ品質問題がデータセット内で検出される場合、データベース内に記憶される根本的原因データセットの一覧上に起こり得るデータ品質問題を有するデータセットの識別子が配置される。出力データ１１０のセットのデータ品質問題が後に検出される場合、出力データ１１０のセットのアップストリームデータ系列要素を識別し、（存在する場合には）それらのアップストリームデータ系列要素の何れが根本的原因データセットの一覧に含まれているかを判定するためにデータベースを照会することができる。

【００５８】

一部の例では、基準データセットの新バージョン又はソースデータセット内等、起こり得るデータ品質問題がデータセット内で検出される場合、ユーザ通知を使用可能にすることができる。一部の例では、データ品質問題を示すために警告フラグを記憶することができる。例えば、起こり得るデータ品質問題が基準データセットの新バージョン内で検出される場合、基準データの新バージョンに関するプロファイルデータと共に警告フラグを記憶することができる。起こり得るデータ品質問題がソースデータセット内で検出される場合、そのソースデータセットに関するプロファイルデータと共に警告フラグを記憶することができる。一部の例では、起こり得るデータ品質問題の存在を示すために警告メッセージをユーザに伝達することができる。警告メッセージは、例えばユーザインタフェース上のメッセージ、アイコン、ポップアップウィンドウ、電子メール、ショートメッセージサービス（ＳＭＳ）メッセージ、又は別の形式とすることができる。

【００５９】

一部の例では、警告フラグ又は警告メッセージが使用される基準プロファイルからの１つ又は複数の限界偏差を規則が指定することができる。例えば、現在のデータセットのプロファイルとそのデータセットに関する基準プロファイルとの間の偏差が、１～２標準偏差等だけ小さい場合に警告フラグを記憶することができ、偏差が２標準偏差を上回る場合に警告メッセージを伝達することができる。限界偏差は、各ソースデータセット及び基準データセットに固有であり得る。

【００６０】

偏差が極度である、例えば基準プロファイルを３標準偏差だけ上回る場合等の一部の例では、ユーザが介入するまでデータ処理システムによる更なる処理を停止することができる。例えば、極度の偏差を有するソースデータ又は基準データの影響を受ける任意の更なる処理が一時停止される。一時停止すべき変換は、影響を受けるソースデータ又は基準データのダウンストリームにあるデータ系列要素を参照するデータによって識別され得る。

【００６１】

一部の例では、基準プロファイルデータが自動で決定される。例えば、所与のデータセットの基準プロファイルデータをそのデータセットの過去のプロファイルデータの履歴的な移動平均として（例えば、そのデータセットの新たなプロファイルデータが決定されるたびに基準プロファイルデータを再計算することによって）自動で更新することができる。一部の例では、ユーザが、例えば、所望の特性を有するデータセットをプロファイリングすることにより、最初の基準プロファイルデータを供給することができる。

【００６２】

変換要素１０６、１０８、１１６のそれぞれに対する最近の更新の時間又は日付等、出力データのデータ系列内の変換要素１０６、１０８、１１６の更新状態を追跡することができる。変換要素に対する最新の更新のタイミングにアクセスすることができ、ユーザは

変換要素の１つ又は複数、例えば正しくない又は破損した変換要素が、出力データ１１０内の既存の又は潜在的なデータ品質問題のあり得る根本的原因かどうかを評価することができる。例えば、出力データ１１０が変換要素１１６から出力される直前に変換要素１１６が更新された場合、出力データ１１０内の既存の又は潜在的なデータ品質問題のあり得る根本的原因として変換要素１１６を識別することができる。

【００６３】

図５を参照すると、追跡エンジン５００は、ソースデータ及び基準データ等のデータ系列要素のプロファイル、並びにデータ処理システムによって生成される出力データ等の所与のデータセットのアップストリームデータ系列内の基準データ及び変換等のデータ系列要素に対する更新をモニタする。

【００６４】

追跡エンジン５００は、データ処理システムによって生成される出力データ等の所与のデータセットのアップストリームにあるデータ系列要素を参照するデータ５０４を記憶する、データ系列リポジトリ５０２を含む。例えば、データ系列リポジトリ５０２は、各データ系列要素の識別子及びデータ系列要素間の関係を示すデータを記憶し得る。データ系列リポジトリ５０２は、ファイル、データベース、又は別のデータ記憶機構であり得る。

【００６５】

追跡エンジン５００は、更新モニタ５０６を含む。更新モニタ５０６は、データ処理システム内の変換要素及び基準データセットが何れの時点で更新されるかをモニタする。データ系列リポジトリ５０２によって参照される変換要素ごとに、更新モニタ５０６は、変換要素を実装するソフトウェアが何れの時点で更新されるかをモニタする。更新が生じるとき、更新モニタ５０６は、ファイル、データベース、別のデータ記憶機構等の更新リポジトリ５０８内にエントリ５１０を記憶する。エントリ５１０は、ソフトウェアが更新された日付、時間、その両方等の更新のタイミングを示す。一部の例では、エントリ５１０が、更新についての手入力された説明、更新によって変更された命令行のテキスト、更新の性質についての別の指示等、更新の性質についての指示も含み得る。更新リポジトリ５０８は、変換要素の識別子により、更新のタイミングにより、又はその両方により索引を付けられ得る。

【００６６】

データ系列リポジトリ５０２によって参照される基準データセットごとに、更新モニタ５０６は、基準データセットが何れの時点で更新されるかをモニタする。更新が生じるとき、更新モニタ５０６は、ファイル、データベース、別のデータ記憶機構等のプロファイルリポジトリ５１６内にエントリ５１４を記憶する。エントリ５１４は、基準データセットが更新された日付、時間、その両方等の更新のタイミングを示す。プロファイルリポジトリ５１６は、基準データセットの識別子により、更新のタイミングにより、又はその両方により索引を付けられ得る。

【００６７】

基準データセットが更新されるとき、その基準データセットに関する品質要素が、基準データの新バージョンと呼ばれる場合もある更新された基準データのプロファイルを生成する。品質要素は、ファイル、データベース、別の記憶機構等の規則リポジトリ５２２内に記憶される重要なデータ要素の一覧５２０に従ってプロファイルを生成することができる。重要なデータ要素とは、ユーザ又はシステムにとって重要であることが分かっているデータ記録内のフィールド、例えばユーザによって指定されるフィールド又は自動で識別されるフィールドである。基準データの新バージョンに関する重要なデータ要素ごとにプロファイルが生成される。例えば、所与の重要なデータ要素について生成されるプロファイルは、重要なデータ要素に関する幾つの個別値が基準データセット内にあるか、及び各個別値が発生する回数を示す統計調査データであり得る。それぞれの重要なデータ要素の生成プロファイルを示す基準プロファイルデータ５２４が、例えば基準データに対する更新を示すエントリ５１４に関連してプロファイルリポジトリ５１６内に記憶される。

【００６８】

ソースデータがデータ処理アプリケーションに与えられるとき、データ系列リポジトリ５０２によって参照される各ソースデータセットのプロファイルが対応する品質要素によって生成される。ソースデータ内の重要なデータ要素ごとにプロファイルが生成され、重要なデータ要素は、規則リポジトリ５２２内に記憶される重要なデータ要素の一覧５２０内で指定される。プロファイルされた各ソースデータセットの生成プロファイルを示すソースプロファイルデータ５２６が、ファイル、データベース、別のデータ記憶機構等のプロファイルリポジトリ５１６内に記憶される。

【００６９】

一部の例では、ダウンストリーム出力データ内でデータ品質問題が生じる場合にのみ、基準プロファイルデータ５２４及びソースプロファイルデータ５２６がアクセスされる。一部の例では、基準データの新バージョン又は受信されたソースデータそれぞれの潜在的なデータ品質問題をデータが示すかどうかを判定するために、基準プロファイルデータ５２４、ソースプロファイルデータ５２６、又はその両方がプロファイルモジュールによって分析される。プロファイルデータ５２４、５２６は、プロファイルの生成直後に分析することができ、又は後の時点において、例えば追跡エンジンが分析のために空いた計算資源を有する任意の時点において分析することができる。

【００７０】

基準プロファイルデータ５２４又はソースプロファイルデータ５２６を分析するために、分析モジュール５３０が、規則リポジトリ５２２内に記憶される自動生成規則又はユーザ指定規則等の規則５３６を適用する。規則は、例えばデータセットごとの１つ又は複数の重要なデータ要素、データ品質問題を引き起こし得る限界偏差、又は別の種類の規則を示し得る。

【００７１】

一部の例では、潜在的なデータ品質問題が基準データの新バージョン内又はソースデータセット内で検出される場合、データ系列リポジトリ５０２内に記憶される根本的原因データセットの一覧５５０上に潜在的なデータ品質問題を有するデータセットの識別子が配置される。ユーザがダウンストリームデータセットのデータ品質問題を後に検出する場合、出力データセットのアップストリームにあるデータ系列要素を識別し、（存在する場合には）それらのアップストリームデータ系列要素の何れが根本的原因データセットの一覧５５０上に含まれているかを識別するために、ユーザは、データ系列リポジトリ５０２を照会することができる。

【００７２】

一部の例では、起こり得るデータ品質問題があるかどうかを判定するために出力データ１１０が自動で分析される。例えば、現在の出力データ１１０のプロファイルを出力データ１１０の旧バージョンの基準プロファイルと比較するために、例えば出力データ１１０の各バッチ又は時間間隔をプロファイルすることができ、プロファイリング規則及び検証規則を出力データ１１０に適用することができる。出力データプロファイリング規則内で指定されるように、現在の出力データ１１０のプロファイルが基準プロファイルから閾値量を上回って外れる場合、潜在的なデータ品質問題を有するものとして現在の出力データ１１０を識別することができる。出力データ検証規則内で指定されるように、現在の出力データ１１０内の特定のデータ要素が期待値域から閾値量を上回って外れる値を有する場合、潜在的なデータ品質問題を有するものとして現在の出力データ１１０を識別することができる。データウェアハウス内に警告フラグを出力データ１１０と共に記憶することができ、又はユーザは、例えば、ユーザインタフェース又はメッセージによって通知され得る。

【００７３】

一部の例では、潜在的なデータ品質問題を有するものとしてユーザが出力データ１１０のセットを識別する。例えば、複数の出力データ１１０のセットを要約するレポートを作成するビジネスアナリストは、自らが分析している他の出力データセットに比べて特定の出力データ１１０のセットが殆ど意味をなさないことを認識する場合がある。アナリスト

は、潜在的なデータ品質問題を有するものとしてその特定の出力データ１１０のセットのフラグを立てることができる。

【００７４】

出力データがデータ品質問題を有する場合、データ品質問題の根本的原因を識別するために、追跡エンジン５００内に記憶されている情報がアクセスされ得る。例えば、ファイル名又はタイムスタンプ等の出力データの識別子が、例えば自動で又はユーザによって照会モジュール５４８に与えられ得る。照会モジュール５４８は、識別された出力データに関連し得る情報を求めて関連リポジトリのそれぞれを照会する。具体的には、照会モジュール５４８は、識別された出力データが依拠する変換、ソースデータ、及び基準データを識別するためにデータ系列リポジトリ５０２を照会する。次いで、照会モジュール５４８は、出力データを処理する直前に生じた識別された変換要素の何れかに対する更新を示す任意のエントリ５１０を求めて更新リポジトリを照会することができる。照会モジュール５４８は、関連する基準プロファイルデータ５２４及び関連する任意の警告フラグと共に、識別された基準データに対する更新を示す任意のエントリ５１４を求めてプロファイルリポジトリ５１６を照会することができる。照会モジュール５４８は、識別された任意のソースデータセットに関するソースプロファイルデータ５２６を求めてプロファイルリポジトリ５１６を照会することができる。

【００７５】

照会モジュール５４８による照会に応じて返される結果がユーザインタフェース上で表示される。ディスプレイは、出力データ内のデータ品質問題の潜在的な根本的原因の理解を得るために、ユーザがデータを見て操作することを可能にする。例えば、出力データが処理される直前に変換要素に対するソフトウェア更新があった場合、更新についての説明又は変更された命令行等、ユーザはその更新に関連する情報を見ることができる。基準プロファイルデータ又はソースプロファイルデータに関連する警告フラグがあった場合、ユーザはプロファイルデータを見ることができる。

【００７６】

一部の例では、照会モジュール５４８によって返される結果は、潜在的なデータ品質問題を有する出力データに関する処理を変換要素が行う直前に変換要素に対する更新が行われたことを示し得る。これを最近更新された変換要素と呼ぶ場合がある。直前とは、例えば、処理の１０分以内、１時間以内、１日以内、別の時間内等、設定された時間内を意味する。更新モニタ５０６は、最近更新された変換要素の１つ又は複数が出力データ内のデータ品質問題の潜在的な根本的原因かどうかを示し得る、最近更新された変換要素に関する追加情報を得ることができる。例えば、更新モニタ５０６は、最近更新された変換要素に関連する任意の処理アーティファクトを識別することができる。処理アーティファクトがあることは、最近更新された変換要素の潜在的な問題を示し得る。最近更新された変換要素に対する更新を更新ログが反映することを確実にするために、更新モニタ５０６は最近更新された変換要素に関連する更新ログを点検することができる。更新ログと最近更新された変換要素に対する更新を示すデータ５１０との間の不一致は、変換要素の潜在的な問題を示し得る。最近更新された変換要素の更新中に取り込まれている可能性がある潜在的な誤りを識別するために、更新モニタ５０６はチェックサム又は他のシステムデータを点検することができる。

【００７７】

一部の例では、最近更新された変換要素の潜在的な問題が検出される場合、ユーザ通知を使用可能にすることができる。一部の例では、潜在的な問題を示すために警告フラグを、例えば更新を示すデータ５１０と共に更新リポジトリ５０８内に記憶することができる。一部の例では、最近更新された変換要素の潜在的な問題の存在を示すために通信モジュール５４６によって警告メッセージをユーザに伝達することができる。例えば、警告メッセージは、ユーザインタフェース上のメッセージ、アイコン、ポップアップウィンドウ、電子メール、ＳＭＳメッセージ、又は別の形式とすることができる。一部の例では、データ系列及びデータ品質の分析を粗粒度のデータ系列と呼ぶ場合があるデータセットのレベ

ルとすることができる。粗粒度のデータ系列は、ダウンストリームデータセットのデータ系列を見る。ダウンストリームデータセットを生成するために使用されるアップストリームデータセット及びアップストリーム変換要素は、ダウンストリームデータセットのデータ系列内にあると見なされる。一部の例では、データ系列及びデータ品質の分析を細粒度のデータ系列と呼ぶ場合がある個々のフィールドのレベルとすることができる。細粒度のデータ系列は、ダウンストリームデータセット内の特定のフィールドのデータ系列を見る。ダウンストリームデータセット内の特定のフィールドを生成するために使用されるアップストリーム変換要素及びアップストリームデータセット内のフィールドは、ダウンストリームデータセットのデータ系列内にあると見なされる。データ品質の分析についてここで説明する手法は、粗粒度のデータ系列及び細粒度のデータ系列の両方に関連して適用され得る。

【００７８】

プロファイリングに関する更なる情報は、参照によりその全内容を本明細書に援用する「Data Profiling」という名称の米国特許第８，８６８，５８０号に見出すことができる。典型的には、データ記録はデータフィールドの組に関連し、各フィールドは各記録に関する特定の値（場合によりヌル値を含む）を有する。一部の例では、データセット内のデータ記録が決まった記録構造を有し、かかる記録構造内では各データ記録が同じフィールドを有する。一部の例では、データセット内のデータ記録が、例えば可変長ベクトル又は条件付きフィールドを含む可変記録構造を有する。一部の例では、プロファイルモジュール２１８が、データセット内のデータ記録に関する初期フォーマット情報をプロファイル要素１５０、１５２、１５４に与えることができる。初期フォーマット情報は、例えば個別値を表すビット数（例えば、１６ビット）、記録フィールドに関連する値及びタグ又は区切り符号に関連する値を含む値の順序、ビットによって表わされる値の種類（例えば、文字列、符号付き／符号なし整数又は他の種類）、又は他のフォーマット情報を含み得る。フォーマット情報は、規則リポジトリ５２２内に記憶されるデータ操作言語（ＤＭＬ）ファイル内で指定され得る。プロファイル要素１５０、１５２、１５４は、ＳＱＬテーブル、ＸＭＬファイル、ＣＳＶファイル等の様々な共通データシステムフォーマットのデータを自動で解釈するために既定のＤＭＬファイルを使用することができ、又は専用のデータシステムフォーマットを記述する、規則リポジトリ２２２から得られるＤＭＬファイルを使用することができる。

【００７９】

図６は、ユーザが出力データセット内の潜在的なデータ品質問題の根本的原因を調査することを可能にするユーザインタフェース３００の一例を示す。ユーザインタフェース３００により、ユーザは出力データセットの識別子３０２又は出力データ内の特定のデータ要素の識別子３０４を入力することができる。例えば、識別子３０２又は３０４は、潜在的なデータ品質問題を有する出力データセット又は特定のデータ要素を識別することができる。図６の例では、ユーザが出力データセット「Billing_records.dat.」を入力している。ユーザインタフェース３００上に、出力データの識別されたセット又は識別されたデータ要素のアップストリームのデータ系列要素をグラフィカルに示す対話型データ系列図３１０が表示される。データ系列図３１０の例では、識別された出力データセットのアップストリームのデータ系列要素がソースデータ３１２、３１４の２つのセット、２つの変換要素３１６、３１８、及び１つの基準データ３２０のセットを含む。

【００８０】

この例のソースデータ３１２、変換要素３１８、基準データ３２０等、起こり得るデータ品質問題を有するアップストリームデータ系列要素が警告フラグ３２４ａ、３２４ｂ、３２４ｃでそれぞれ印付けされる。ユーザは、警告フラグ上でクリックし若しくはタップし、警告フラグ上にマウスポインタを重ね、又は他の方法で警告フラグを選択すること等によって警告フラグを選択し、関連する起こり得るデータ品質問題に関する情報にアクセスすることができる。データセットに関連する起こり得るデータ品質問題に関する情報は、プロファイルデータ、１つ又は複数のデータ要素に関する基準プロファイルデータ、プ

ロファイルデータの統計的分析の結果（基準プロファイルデータからのプロファイルデータの偏差等）、検証規則によって指定された許容値を満たさない値、又は他の情報を含み得る。変換要素に関連する起こり得るデータ品質問題に関する情報は、変換要素に対する直近の更新日、更新についての説明、更新に由来する符号の抜粋、又は他の情報を含み得る。一部の例では、警告フラグの１つをユーザが選択することに応じてデータ系列図上に情報バブルをオーバレイすることができる。一部の例では、警告フラグの１つをユーザが選択することに応じて新たな画面を表示することができる。一部の例では、情報バブル又は新たな画面内に表示される情報を対話型とすることができ、それにより、ユーザは、情報片を選択することによって更なる詳細情報にアクセスすることができる。

【００８１】

ユーザインタフェース３００により、ユーザは規則エディタ３２８にアクセスすることもでき、規則エディタ３２８によってユーザはプロファイリング規則、検証規則、又はその両方を追加し、削除し、又は修正することができる。例えば、ユーザは、データセットごとに重要なデータ要素を追加し、削除し、又は修正すること、潜在的なデータ品質問題の識別を引き起こす限界偏差を更新すること、プロファイリング規則又は検証規則を新たなデータセットの受信時に自動で適用すべきかダウンストリームのデータ品質問題を検出したときにのみ自動で適用すべきかを指定すること、又はプロファイリング規則若しくは検証規則に対して他の変更を加えることができる。

【００８２】

ある具体例では、データ処理システムが通話記録を処理して課金記録を生成する。各ソースデータ記録は通話を表し、通話の日付、時間、通話の持続時間、ダイヤル側の電話番号、着呼側の電話番号等のデータを記憶するフィールドを含む。請求書を生成するために、ソースデータ記録がバッチ処理内で毎月処理される。この例では、２０１５年５月の月に顧客アカウントの９５％について請求書が生成されなかった。ユーザは、２０１５年５月の請求書を生成するために使用された出力データのアップストリームデータ系列内のデータ系列要素のプロファイル及びかかるデータ系列要素に対する更新に関する情報を要求した。基準ソースプロファイルデータがダイヤル側電話番号フィールド内の１５０万個～２４０万個の固有値の期待範囲を示した一方、ソースプロファイルデータは、２０１５年５月の請求書を生成するために使用されたソースデータ記録内のダイヤル側電話番号フィールドが１０個の固有値のみを有したことを示した。ソースプロファイルデータのこの点検に基づき、ユーザはソースデータ記録が破損していると判定した。２０１５年５月の請求書を正しく生成するためにソースデータ記録が圧縮記憶域から取得され、再び処理された。

【００８３】

別の具体例では、データ処理システムが企業内財務記録を処理し、各財務記録を企業の事業部に割り振る。企業の事業部に対する各財務記録の割り振りは、各記録内の部門識別子を企業基準データセットによって提供される企業の６つの事業部のうちの１つにマッピングすることによって行われる。企業基準データに関する基準プロファイルデータは、企業の事業部の数が過去１０年にわたって一貫して６つであったことを示した。基準データは毎四半期更新される。直近の更新後、基準データがプロファイルされ、基準データ内の企業の事業部の数が６０まで増えたことを示した。６つの事業部の基準からの更新された基準データのプロファイルの偏差は、システム管理者への警告メッセージの送信を引き起こすのに十分大きかった。加えて、基準データを調査し、必要に応じて訂正できるまでデータ処理システムによる更なる処理が一時停止された。

【００８４】

図７を参照すると、ある具体例では、データ処理システム５０が、２０１６年４月１日にthebostonshop.comで行われたオンライン購入の記録を含む入力データ５８を処理する複数の変換要素５２、５４、５６を含む。入力データ５８の各記録は、州フィールドを含む複数のフィールドを含む。この例では、コンポーネント５６は、入力データの州フィールド内の値に基づいて８つのファイル６０ａ～６０ｈのうちの１つに各データ記録を送信

する分割コンポーネントである。例えば、州フィールド内にＭＡの値を有する記録はファイル６０ａに送信され、値ＴＸを有する記録はファイル６０ｂに送信され、値ＣＡを有する記録はファイル６０ｃに送信され、値ＤＥを有する記録はファイル６０ｄに送信され、値ＮＹを有する記録はファイル６０ｅに送信され、値ＩＬを有する記録はファイル６０ｆに送信され、値ＲＩを有する記録はファイル６０ｇに送信され、他の任意の値を有する記録はファイル６０ｈに送信される。各ファイルに送信される記録の数が図７に示されている。図７の例では、各ファイルに送信される記録の数が期待範囲内にあり、従ってデータ品質の警告は生成されない。これは入力データ５８が期待範囲に含まれるためである。

【００８５】

入力データ５８の品質が品質要素６２によって特徴付けられる。品質要素６２は入力データ５８の州フィールドのプロファイルを生成し、州フィールドのプロファイルと入力データの州フィールドの基準プロファイルとの間の許容偏差を示す自動生成規則を適用する。基準プロファイルは、過去１年にわたってデータ処理システム５０によって処理されたデータの平均プロファイルを表し、それを超えると潜在的なデータ品質問題が識別される許容偏差を示す。この例では、入力データ５８のプロファイル内の州フィールド内の値分布が基準プロファイル内の値分布と１０％を超えて異なる場合、潜在的なデータ品質問題を有するものとして入力データ５８を識別すべきであると自動生成規則が示し、州フィールドの基準プロファイルは。１０％の許容偏差と共に州フィールド内の以下の値分布：

ＭＡ：６％
ＴＸ：２５％
ＣＡ：３３％
ＤＥ：３％
ＮＹ：１７％
ＩＬ：１１％
ＲＩ：４％
他の任意の値：１％

を示し、図７を見れば分かるように、州フィールドの実際のプロファイルが基準プロファイルの１０％の許容偏差に含まれており、従って入力データのデータ品質問題はない。

【００８６】

図８Ａを参照すると、データ処理システム５０の異常動作の一例では、入力データ５５が、２０１６年４月２日にbostonshop.comで行われたオンライン購入の記録を含む。この例では、ファイル６０ｇに記録が送信されていない。データ処理システム５０のオペレータはファイル６０ｇが空であることに気付く場合があり、又は空のファイルがダウンストリームのデータ処理システムによる更なる処理内で誤りを引き起こす可能性がある。データ処理システム５０のオペレータは、ファイル６０ａ～６０ｈのデータ系列内のアップストリームデータ要素の品質を調査することにより、ファイル６０ｇに記録が送信されなかった根本的原因を追跡することができる。具体的には、入力データ５５はファイル６０ａ～６０ｈのアップストリームデータ系列に属する。

【００８７】

図８Ｂも参照すると、品質要素６２は入力データ５５の州フィールドの以下の実際のプロファイルを生成する：

ＭＡ：６％
ＴＸ：２５．１％
ＣＡ：３２．７％
ＤＥ：２．９％
ＮＹ：１７．１％
ＩＬ：１１．１％
ＲＩ：０％
他の任意の値：５．１％。

【００８８】

入力データ５５の州フィールドのプロファイルと州フィールドの基準プロファイルとの間の偏差により、入力データ５５は潜在的なデータ品質問題を有するものとして識別され、潜在的なデータ品質問題を示すために警告フラグが記憶される。オペレータが空のファイル６０ｇの根本的原因を追跡するとき、オペレータは潜在的なデータ品質問題が入力データ５５内に存在したことを容易に認めることができる。例えば、入力データ５５が破損していたかどうか、アップストリームのデータ処理システム内での入力データ５５のそれまでの処理が偏差を引き起こしたかどうか、又は別の原因を判定するために、オペレータはその知識を使用して偏差の原因を調査することができる。例えば、図８Ｃも参照すると、この例では実際の入力データ５５の一部を見ることにより、オペレータは、値「ＲＩ」内の文字が「ＩＲ」と読めるように逆になっており、それらの記録がファイル６０ｇ内ではなくファイル６０ｈ内に記憶されることを引き起こしていると認識し得る。

【００８９】

図９Ａを参照すると、データ処理システム５０の異常動作の別の例では、入力データ６４が、２０１６年４月３日にthebostonshop.comで行われたオンライン購入の記録を含む。この例では、記録がファイル６０ａのみに送信されており、他のファイル６０ｂ～６０ｈの何れにも送信されていない。データ処理システム５０のオペレータは、ファイル６０ｂ～６０ｈが空であることに気付く場合があり、又は空のファイルがダウンストリームのデータ処理システムによる更なる処理内で誤りを引き起こす可能性がある。

【００９０】

図９Ｂも参照すると、データ処理システムのオペレータは、ファイル６０ａ～６０ｈのデータ系列内のアップストリームデータ要素の品質を調査することにより、記録の全てがファイル６０ａに送信された根本的原因を追跡することができる。この例では、品質要素６２が入力データ６４の州フィールドの以下のプロファイルを生成する：

ＭＡ：６．１％
ＴＸ：２５．２％
ＣＡ：３２．６％
ＤＥ：２．９％
ＮＹ：１７．０％
ＩＬ：１１．１％
ＲＩ：４．１％
他の任意の値：１％。

【００９１】

入力データ６４の州フィールドのプロファイルは州フィールドの基準プロファイルと一貫性があり、従って潜在的なデータ品質問題は識別されない。次いで、オペレータは、ファイル６０ａ～６０ｈのデータ系列内にある変換要素５２、５４、５６の更新状態を調査することができる。例えば、オペレータは、入力データ６４を処理する直前に変換要素５６が更新されたと判定することができ、従って変換要素５６が空のファイル６０ｂ～６０ｈの根本的原因であり得る。

【００９２】

図１０Ａを参照すると、ある具体例では、データ処理システム８０が、特定の塔によって扱われる携帯電話の通話に関する通話記録を含む入力データ８６のストリームを処理する複数の変換要素８２、８４を含む。入力データ８６の各記録は、電話＿番号フィールドを含む複数のフィールドを含む。入力データ８６は変換要素８２によってフォーマットされ、その後、電話＿番号フィールド内の値により変換要素８４によってソートされ、待ち行列８８内に出力され、そこから更なる処理のために第２のデータ処理システム９０内に供給される。この例では、待ち行列８８から第２のデータ処理システム９０内に供給される記録の２５％が処理誤りを引き起こす。データ処理システム８０のオペレータは、待ち行列８８のデータ系列内のアップストリームデータ要素の品質を調査することにより、これらの処理誤りの根本的原因を追跡することができる。

【００９３】

入力データ８６の品質が品質要素９０によって特徴付けられ、フォーマット変換要素８２から出力されるデータ９４の品質が品質要素９２によって特徴付けられる。品質要素９０、９２の両方は、電話＿番号フィールド内の値が１０桁の整数であるべきこと、及び記録の３％超が規則を満たさない場合、潜在的なデータ品質問題を識別すべきことを指定するユーザ生成規則を適用する。この例では、品質要素９０は、データ８６内の記録の０．１％が電話＿番号フィールド内で１１桁の整数を有すると判定する。この記録のパーセンテージは３％の閾値を下回るため、品質要素９０は入力データ８６のいかなる潜在的なデータ品質問題も識別しない。品質要素９２は、電話＿番号フィールド内に英数字値を有するものとしてデータ９４内の記録の２５％を特徴付ける。データ９４の一部の一例を図１０Ｂに示す。データ９４の潜在的なデータ品質問題を示すために警告フラグが記憶される。オペレータが処理誤りの根本的原因を追跡するとき、オペレータは、入力データ８６内のデータ品質問題が識別されていないが、潜在的なデータ品質問題がデータ９４内にあることを容易に認めることができる。

【００９４】

図１１を参照すると、ソースデータセットの品質を決定するためのプロセスの一例では、ソースデータセットがデータ処理アプリケーション内に受信される（４００）。ソースデータセットのプロファイルが生成され記憶される（４０２）。ソースデータセットに関する１つ又は複数の規則が取得される（４０４）。ソースデータ又はソースデータのプロファイルが１つ又は複数の規則に従って分析される（４０６）。１つ又は複数の規則がソースデータセットによって満たされない場合（４０８）、潜在的なデータ品質問題を示す警告がプロファイルデータと共に記憶され、ユーザに伝達され、又はその両方が行われ（４１０）、起こり得るデータ品質問題を有するデータセットの一覧にソースデータが追加される。１つ又は複数の規則がソースデータによって満たされる場合（４０８）、ソースデータがデータ処理アプリケーションによって処理される（４１２）。規則によって指定される閾値又は許容値からの極度の偏差等に関する一部の事例では、ユーザによる介入が処理の再開を可能にするまで処理が一時停止される。処理中又は処理後、ユーザは、例えば、ダウンストリームのデータ品質問題の潜在的な根本的原因を調査するために、記憶されたプロファイルデータにアクセスする権限を与えられる。

【００９５】

図１２を参照すると、データ処理システム内の基準データの品質をモニタリングするためのプロセスの一例では、基準データセットがモニタされる（５００）。基準データセットが更新されると、基準データの新バージョンのプロファイルが生成され記憶される（５０２）。例えば、プロファイルの生成は、基準データに対する予定された各更新後に行われ得る。基準データセットに関する１つ又は複数の規則が取得される（５０４）。基準データの新バージョン又は基準データの新バージョンのプロファイルが１つ又は複数の規則に従って分析される（５０６）。１つ又は複数の規則が基準データの新バージョンによって満たされない場合（５０８）、起こり得るデータ品質問題を示す警告がプロファイルデータと共に記憶され、ユーザに伝達され、又はその両方が行われる（５１０）。１つ又は複数の規則が基準データの新バージョンによって満たされる場合（５０８）、データ処理システムによるその後の処理の開始又は続行が許可される（５１２）。規則によって指定される閾値又は許容値からの極度の偏差等に関する一部の事例では、ユーザによる介入が処理の開始又は続行を可能にするまで処理が一時停止される。処理中又は処理後、ユーザは、例えば、ダウンストリームのデータ品質問題の潜在的な根本的原因を調査するために、記憶されたプロファイルデータにアクセスする権限を与えられる。

【００９６】

一部の例では、例えば規則ごとに更新日を決定するために、規則を適用する前に規則が分析される。規則が限界有効期間よりも古い場合、その規則は適用されない場合があり、又はその規則を更新する準備が整っている可能性があることをユーザに警告することができる。

【００９７】

図１３を参照すると、変換要素に対する更新を分析するためのプロセスの一例では、変換要素に対する最近の更新の時間が識別される（６００）。例えば、最近の更新のタイムスタンプをデータリポジトリ内に記憶することができる。変換要素が最近の更新を有さない場合（６０２）、変換要素に対する更新が更に分析されることはない（６０４）。最近の更新は、１０分以内、１時間以内、１日以内、別の時間内等、ある限界時間内の更新とすることができる。変換要素が最近更新された場合（６０２）、任意の処理アーティファクトが識別される（６０６）。更新ログとデータリポジトリ内に記憶された最近の更新のタイムスタンプとの間の任意の不整合を識別するために、変換要素に関連する更新ログが点検される（６０８）。変換要素の更新中に取り込まれている可能性がある任意の潜在的な誤りの指示のために、変換要素に関連するチェックサム又は他のシステムデータが点検される（６１０）。潜在的な問題が識別されない場合（６１２）、システムによる処理の開始又は続行が許可される（６１４）。１つ又は複数の潜在的な問題が識別される場合（６１２）、変換要素の潜在的な問題を示す警告がデータリポジトリ内に記憶され、ユーザに伝達され、又はその両方が行われる（６１６）。データ処理システムによる処理は、開始すること若しくは続行することを認められてもよく、又はユーザによる介入が処理の開始又は続行を可能にするまで一時停止してもよい。

【００９８】

図１４は、プロセスの一例のフローチャートである。データ処理システムによって生成される出力データセットを示す情報が受信される（７００）。出力データセットに関係するデータ系列情報に基づき、出力データセットが依拠する１つ又は複数のアップストリームデータセットが識別される（７０２）。データ系列情報は、出力データセットが依拠する１つ又は複数のデータセット、出力データセットに依拠する１つ又は複数のデータセット、又はその両方を示す。１つ又は複数のデータセットの何れが誤り又は起こり得る誤りを有するかを判定することを含め、データセットの部分集合を識別するために出力データセットが依拠する識別されたアップストリームデータセットのそれぞれを分析する（７０４）。特定のアップストリームデータセットごとに、特定のアップストリームデータセットのプロファイルと特定のアップストリームデータセットに関する基準プロファイルとの間の許容偏差を示す第１の規則が適用され（７０６）、特定のアップストリームデータセット内の１つ又は複数のデータ要素の許容値又は禁止値を示す第２の規則が適用される（７０８）。一部の例では、第１の規則又は第２の規則のみが適用される。第１の規則、第２の規則、又はその両方は自動で生成され又はユーザによって指定され得る。第１の規則、第２の規則、又はその両方を適用した結果に基づき、アップストリームデータセットの１つ又は複数が部分集合のために選択される（７１０）。アップストリームデータセットの部分集合に関連する情報が出力される（７１２）。

【００９９】

図１５は、プロセスの一例のフローチャートである。データ処理システムのダウンストリームデータセットのデータ要素内の誤り又は起こり得る誤りが、例えば自動で又はユーザ入力に基づいて識別される（９００）。ダウンストリームデータセットに関係するデータ系列情報に基づき、データ要素に影響を及ぼす１つ又は複数のアップストリームデータセットが自動で識別される（９０２）。識別されたアップストリームデータセットのそれぞれの現在のプロファイル及び基準プロファイルを分析することを含め、何れのアップストリームデータセットが誤りを有するか又は誤りを有する可能性が高いかを判定する（９０４）。例えば、各アップストリームデータセットは、現在のプロファイルのそれぞれに１つ又は複数の規則を適用することによって分析され得る。規則は、特定のアップストリームデータセットの現在のプロファイルと特定のアップストリームデータセットの対応する基準プロファイルとの間の許容偏差を示し得る。規則は、特定のアップストリームデータセット内のデータ要素に関する許容値を示し得る。誤りを有するか又は誤りを有する可能性が高いアップストリームデータセットのそれぞれに関連する情報が出力される（９０６）。

【０１００】

本明細書に記載のデータ品質をモニタし追跡するための技法は、コンピュータ技術に基づいており、コンピュータによって実装されるプロセスの実行中に生じる問題に対処するために使用することができる。例えば、本明細書に記載のモニタリング及び追跡のための技法を使用し、コンピュータによって実装されるデータ処理システムによるデータセットの処理がモニタされ、より効率的、効果的、又は正確にされ得る。加えて、本明細書に記載の技法は、システム管理者等のユーザがデータ処理システムの動作を管理することを補助するために適用され得る。

【０１０１】

図１６は、モニタリング及び追跡のための技法が使用され得るデータ処理システム１０００の一例を示す。システム１０００はデータソース１００２を含み、データソース１００２は、そのそれぞれが様々なフォーマット（例えば、データベーステーブル、スプレッドシートファイル、フラットテキストファイル、又はメインフレームコンピュータによって使用されるネイティブフォーマット）の何れかによってデータを記憶し又は提供することができる、記憶装置又はオンラインデータストリームへの接続等の１つ又は複数のデータソースを含み得る。データは、ロジスティックデータ、分析データ、又はマシンデータとすることができる。実行環境１００４は、前処理モジュール１００６及び実行モジュール１０１２を含む。実行環境１００４は、例えばＵＮＩＸオペレーティングシステムのバージョン等の適切なオペレーティングシステムの制御下で１つ又は複数の汎用コンピュータ上にホストされ得る。例えば、実行環境１００４は、ローカルである（例えば、対称型マルチプロセッシング（ＳＭＰ）コンピュータ等のマルチプロセッサシステム）若しくはローカル分散された（例えば、クラスタとして結合される複数のプロセッサ若しくは超並列処理（ＭＰＰ）システム、又は遠隔である若しくは遠隔分散された（例えば、ローカルエリアネットワーク（ＬＡＮ）及び／又は広域ネットワーク（ＷＡＮ）によって結合される複数のプロセッサ）、又はその任意の組合せである、複数の中央処理装置（ＣＰＵ）又はプロセッサコアを使用するコンピュータシステムの構成を含むマルチノード並列計算環境を含み得る。

【０１０２】

データソース１００２を提供する記憶装置は、例えば実行環境１００４をホストするコンピュータに接続される記憶媒体（例えば、ハードドライブ１００８）上に記憶され実行環境１００４にとってローカルである場合があり、又は例えば実行環境１００４をホストするコンピュータと（例えばクラウドコンピューティングインフラによって提供される）遠隔接続上で通信する遠隔システム（例えば、メインフレームコンピュータ１０１０）上にホストされ実行環境１００４にとって遠隔的である場合もある。

【０１０３】

前処理モジュール１００６は、データソース１００２からデータを読み取り、実行用のデータ処理アプリケーションを作成する。例えば、前処理モジュール１００６はデータ処理アプリケーションをコンパイルし、実行環境１００４にとってアクセス可能なデータ記憶システム１０１６との間でコンパイルされたデータ処理アプリケーションを記憶及び／又はロードし、実行用のデータ処理アプリケーションを作成するための他のタスクを実行することができる。

【０１０４】

実行モジュール１０１２は、前処理モジュール１００６によって作成されたデータ処理アプリケーションを実行してデータセットを処理し、その処理から生じる出力データ１０１４を生成する。出力データ１０１４はデータソース１００２内に再び記憶することができ、実行環境１００４にとってアクセス可能なデータ記憶システム１０１６内に記憶することができ、又は他の方法で使用され得る。データ記憶システム１０１６は、実行モジュール１０１２によって実行されるデータ処理アプリケーションを開発者１０２０が設計し編集することができる開発環境１０１８にとってもアクセス可能である。一部の実装形態では、開発環境１０１８は頂点間の（作業要素、即ちデータのフローを表す）有向辺によってつながれる（データ処理コンポーネント又はデータセットを表す）頂点を含むデータ

フローグラフとしてアプリケーションを開発するためのシステムである。例えば、かかる環境は、参照により本明細書に援用する「Managing Parameters for Graph-Based Applications」という名称の米国特許出願公開第２００７／００１１６６８号でより詳細に説明されている。かかるグラフベースの計算を実行するためのシステムは、参照によりその全内容を本明細書に援用する「EXECUTING COMPUTATIONS EXPRESSED AS GRAPHS」という名称の米国特許第５，９６６，０７２号で説明されている。このシステムに従って作成されるデータフローグラフは、グラフコンポーネントによって表わされる個々のプロセスの内外に情報を出し入れするための、プロセス間で情報を移動させるための、及びプロセスの実行順序を定めるための方法を提供する。このシステムは、使用可能な任意の方法からプロセス間通信方法を選択するアルゴリズムを含む（例えば、グラフのリンクによる通信経路は、プロセス間でデータを渡すためにＴＣＰ／ＩＰ又はＵＮＩＸドメインソケットを使用することができ、又は共用メモリを使用する）。

【０１０５】

前処理モジュール１００６は、様々な形式のデータベースシステムを含むデータソース１００２を具体化し得る様々な種類のシステムからデータを受信することができる。データは、場合によりヌル値を含むそれぞれのフィールド（「属性」又は「カラム」とも呼ばれる）のための値を有する記録として編成され得る。データソースからデータを最初に読み取るとき、前処理モジュール１００６は、典型的には、そのデータソース内の記録に関する何らかの初期フォーマット情報から始める。一部の状況では、データソースの記録構造が最初に分かっていない場合があり、代わりにデータソース又はデータの分析後に決定され得る。記録に関する初期情報は、例えば個別値を表すビット数、記録内のフィールドの順序、及びビットによって表される値の種類（例えば、文字列、符号付き／符号なし整数）を含み得る。

【０１０６】

上記のモニタリング及び追跡の手法は、適切なソフトウェアを実行する計算システムを使用して実装され得る。例えば、ソフトウェアは、１つ又は複数のプログラムされた又はプログラム可能な計算システム（分散、クライアント／サーバ、グリッド等の様々なアーキテクチャのものとすることができる）上で実行される１つ又は複数のコンピュータプログラム内の手続きを含むことができ、かかる計算システムは、少なくとも１個のプロセッサ、少なくとも１つのデータ記憶システム（揮発性及び／又は不揮発性のメモリ及び／又は記憶素子を含む）、（少なくとも１つの入力装置又はポートを使用して入力を受け付けるための、及び少なくとも１つの出力装置又はポートを使用して出力を与えるための）少なくとも１つのユーザインタフェースをそれぞれ含む。ソフトウェアは、例えばグラフの設計、構成、及び実行に関係するサービスを提供するより大きいプログラムの１つ又は複数のモジュールを含み得る。プログラムのモジュール（例えば、グラフの要素）は、データリポジトリ内に記憶されるデータモデルに適合するデータ構造又は他の編成されたデータとして実装され得る。

【０１０７】

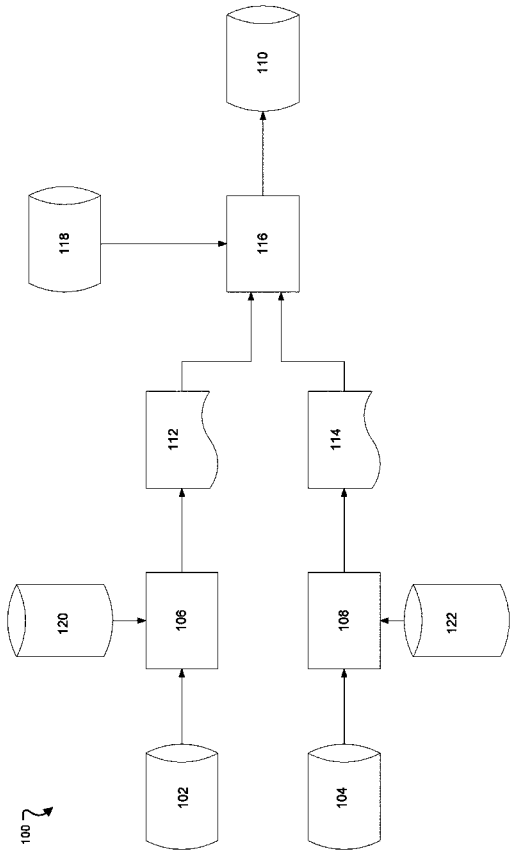
ソフトウェアは、（例えば、汎用又は専用の計算システム又は装置によって読取可能な）ＣＤ－ＲＯＭ又は他のコンピュータ可読媒体等の有形の非一時的媒体上に与えることができ、又はソフトウェアの実行場所である計算システムの有形の非一時的媒体にネットワークの通信媒体上で運ぶ（例えば、伝搬信号内に符号化する）ことができる。処理の一部又は全てを専用コンピュータ上で、又はコプロセッサ、書替え可能ゲートアレイ（ＦＰＧＡ）、専用の特定用途向け集積回路（ＡＳＩＣ）等の専用ハードウェアを使用して実行することができる。処理は、ソフトウェアによって指定される計算の様々な部分が異なる計算要素によって実行される分散方式で実装されてもよい。そのような各コンピュータプログラムは、本明細書に記載の処理を実行するために記憶装置媒体がコンピュータによって読み取られるとき、コンピュータを構成し操作するために、好ましくは汎用又は専用のプログラム可能コンピュータによってアクセス可能な記憶装置のコンピュータ可読記憶媒体（例えば、ソリッドステートメモリ若しくは媒体、又は磁気若しくは光学媒体）上に記憶

され又はかかるコンピュータ可読記憶媒体にダウンロードされる。本発明のシステムは、コンピュータプログラムで構成される有形の非一時的媒体として実装されると考えることもでき、そのように構成される媒体は、本明細書に記載の処理ステップの１つ又は複数を実行するためにコンピュータを特定の且つ既定の方法で動作させる。

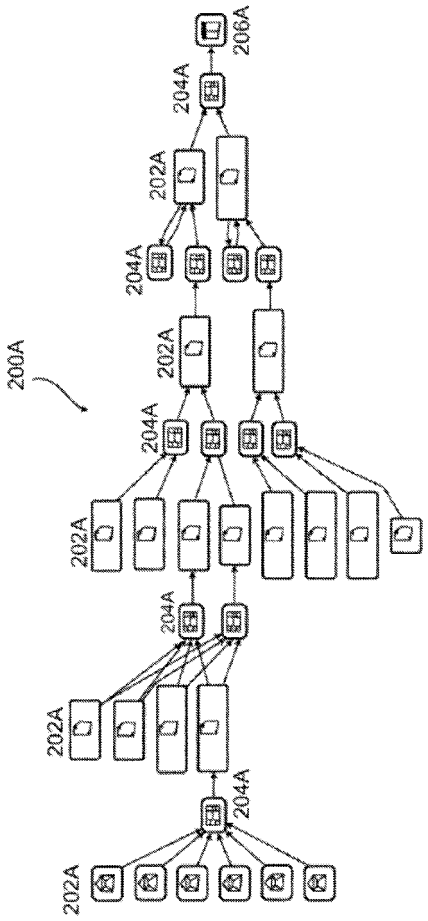
【０１０８】

本発明の幾つかの実施形態について説明してきた。それでもなお、上記の説明は、添付の特許請求の範囲によって定める本発明の範囲を限定するのではなく、例示を目的とすることを理解すべきである。従って、他の実施形態も添付の特許請求の範囲に含まれる。例えば、本発明の範囲から逸脱することなく様々な修正形態がなされ得る。加えて、上記のステップの一部は順序に左右されない場合があり、従って記載したのと異なる順序で実行することができる。

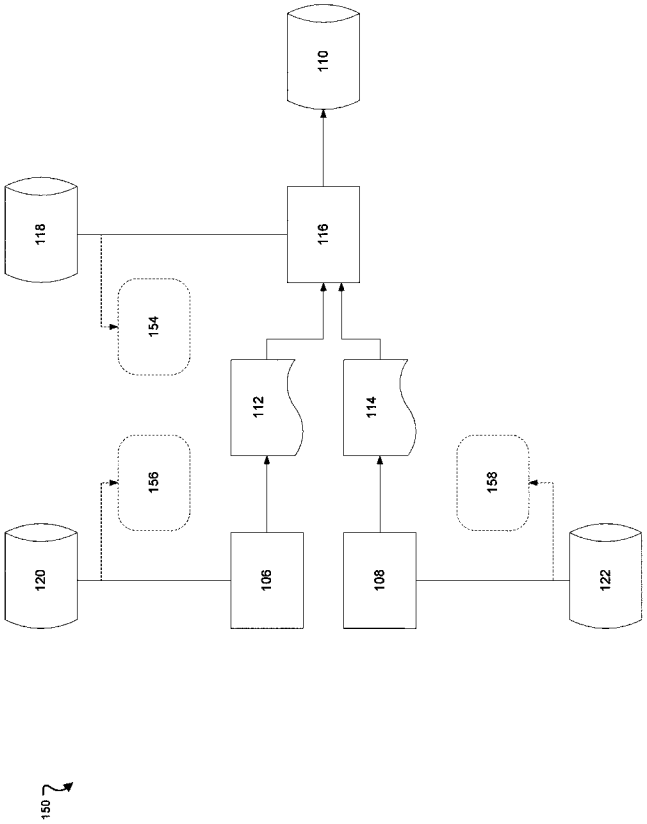
【図１】



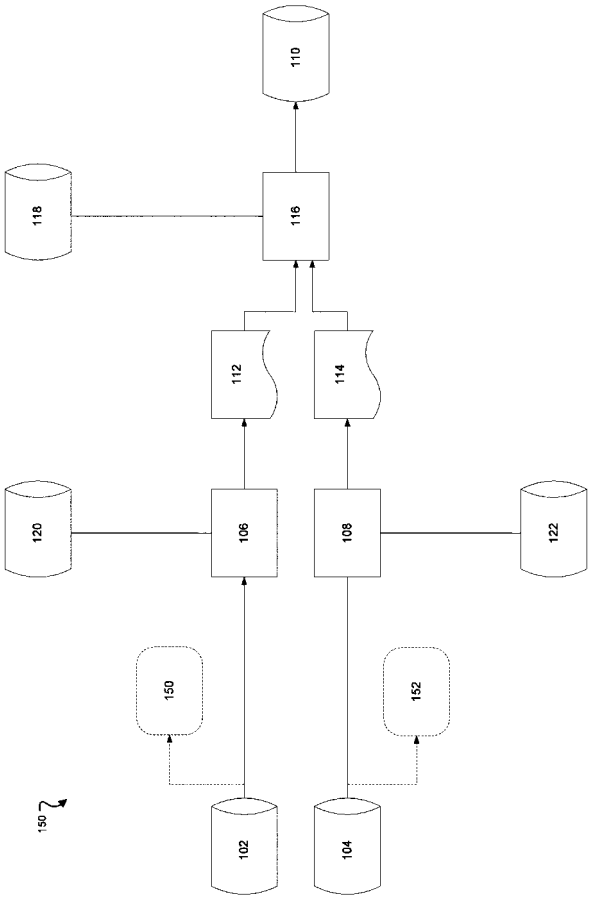
【図２】



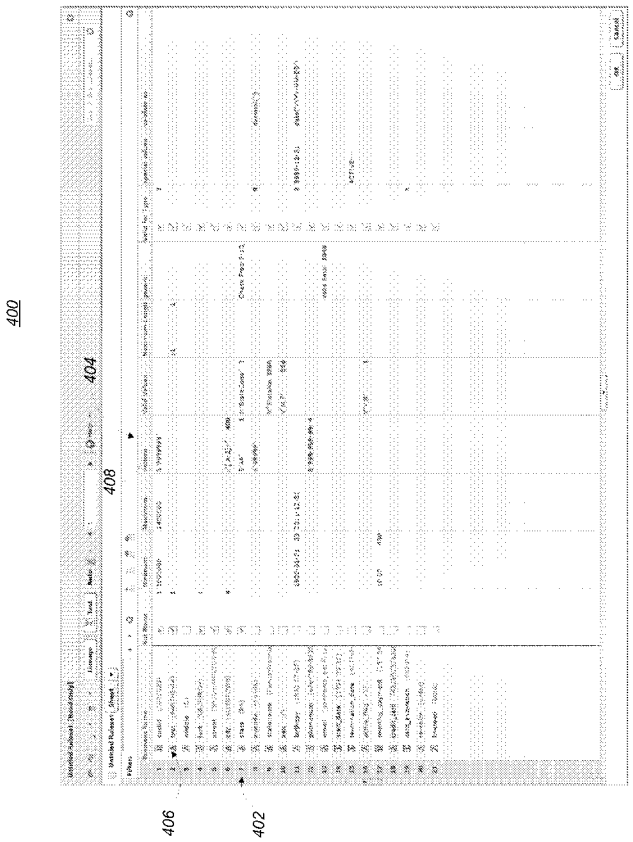
【図３Ａ】


110
118
116
154
112
114
156
158
120
106
108
122
150

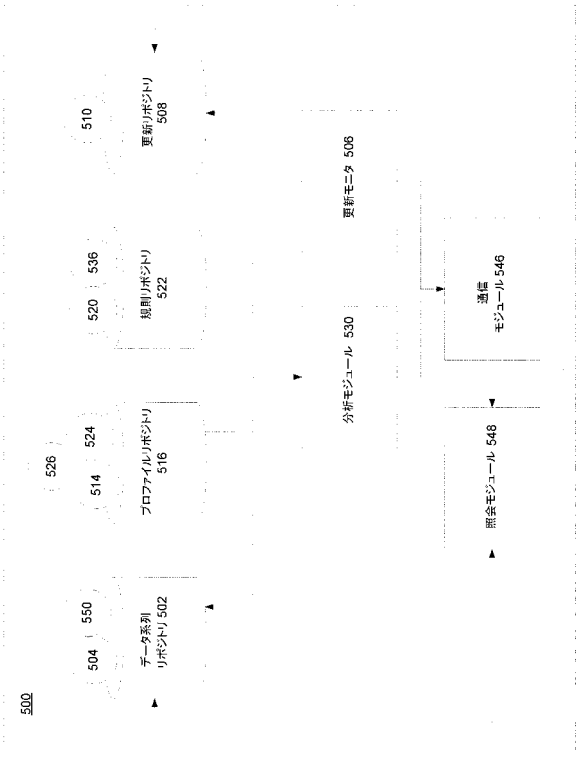
【図３Ｂ】


110
118
116
112
114
120
106
108
122
150
152
102
104
150

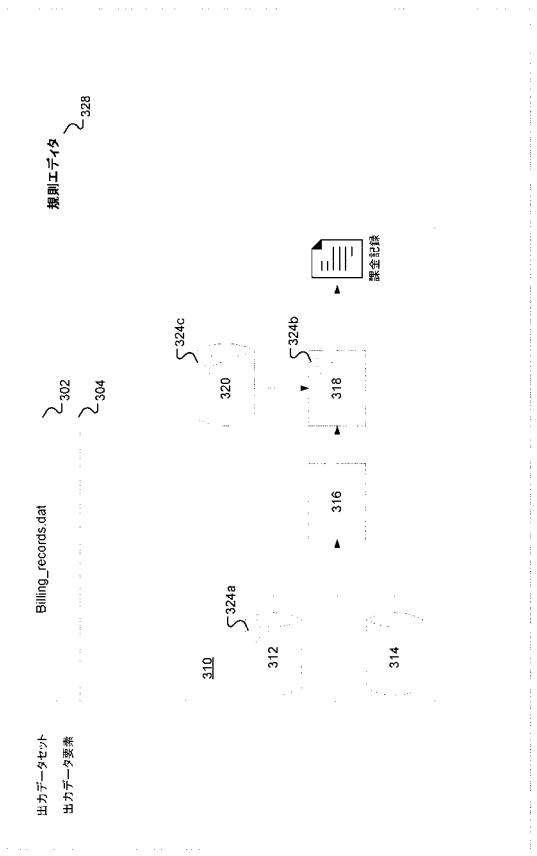
【図４】


400
404
408
406
402

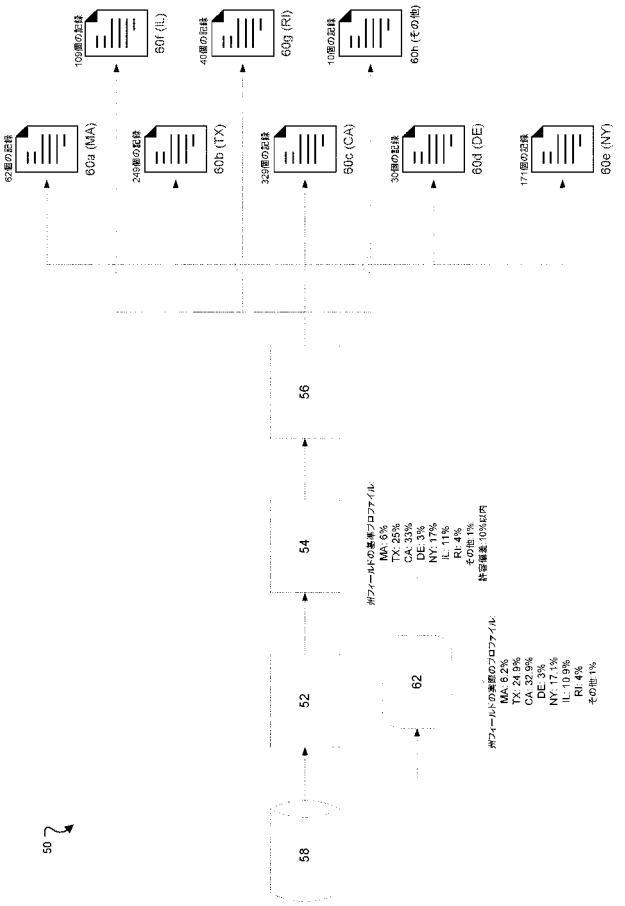
【図５】


500
504
550
データ系列
リポジトリ 502
526
514
524
プロファイルリポジトリ
516
520
536
規則リポジトリ
522
510
更新リポジトリ
508
分析モジュール 530
更新モニタ 506
通信
モジュール 546
照会モジュール 548

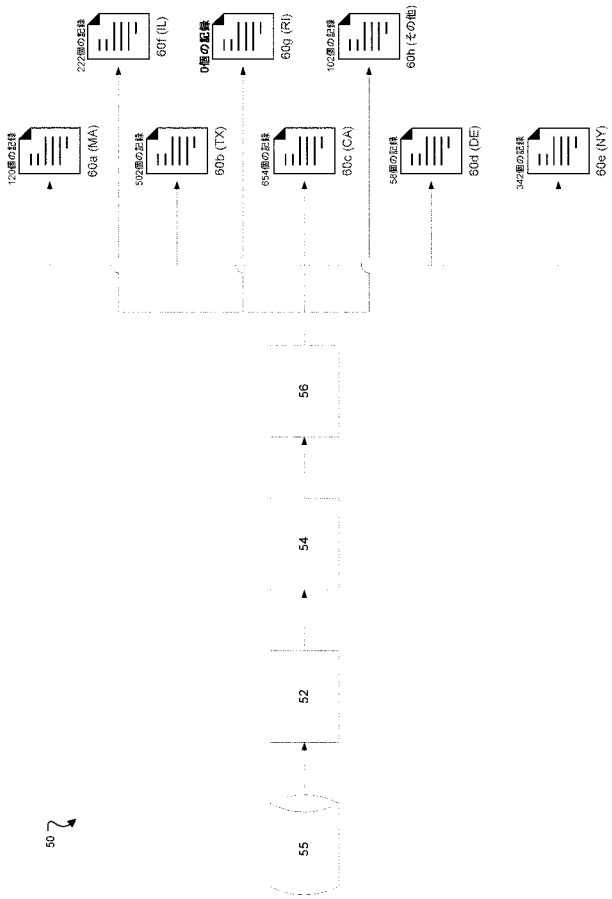
【図６】


規則エディタ
328
課金記録
324c
324b
320
318
302
304
316
Billing_records.dat
324a
310
312
314
出力データセット
出力データ要素

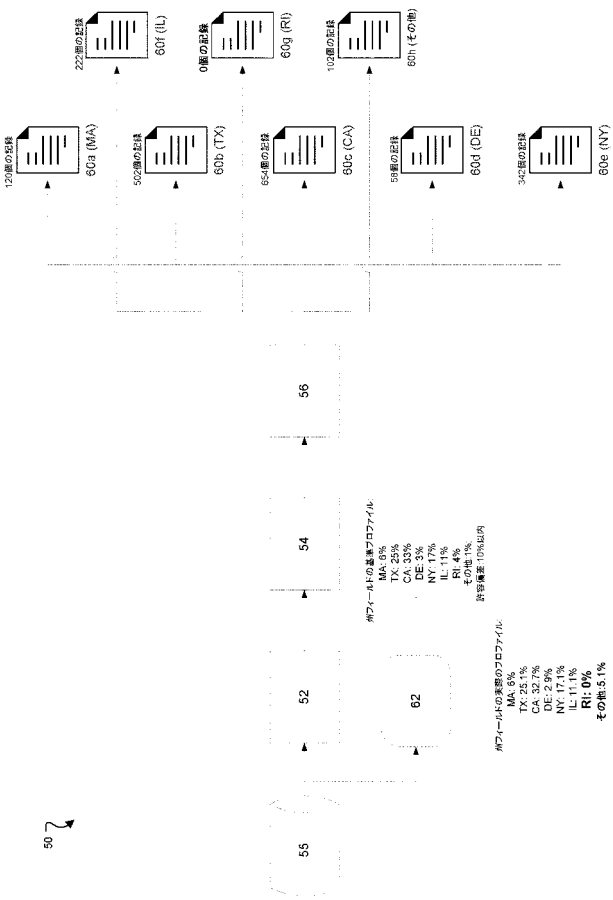
【図７】


60a (MA)
60b (TX)
60c (CA)
60d (DE)
60e (NY)
60f (IL)
60g (RI)
60h (その他)
56
54
52
62
58
50

【図８Ａ】


60a (MA)
60b (TX)
60c (CA)
60d (DE)
60e (NY)
60f (IL)
60g (RI)
60h (その他)
56
54
52
55
50

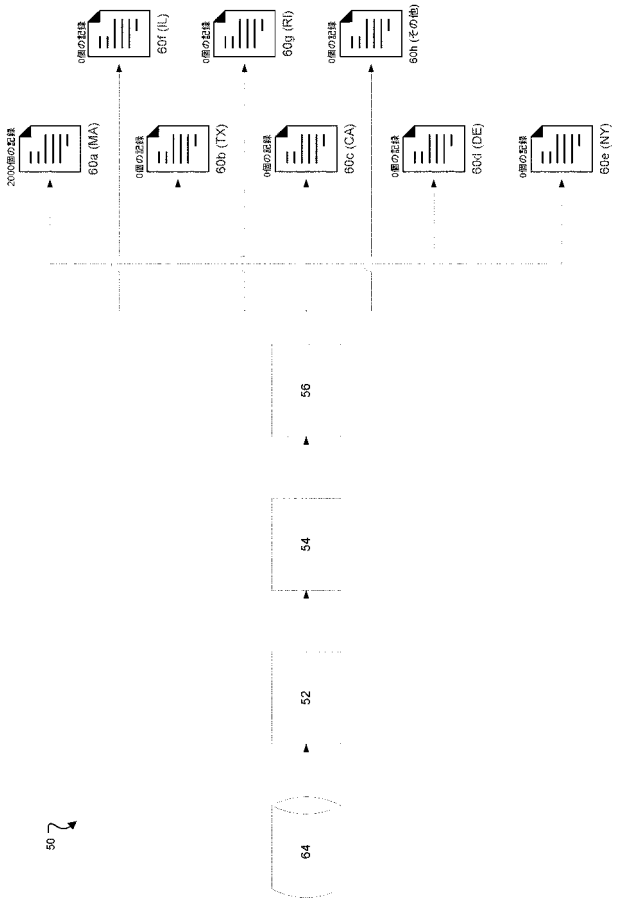
【図８Ｂ】


60a (MA)
60b (TX)
60c (CA)
60d (DE)
60e (NY)
60f (IL)
60g (RI)
60h (その他)
56
54
52
62
55
50

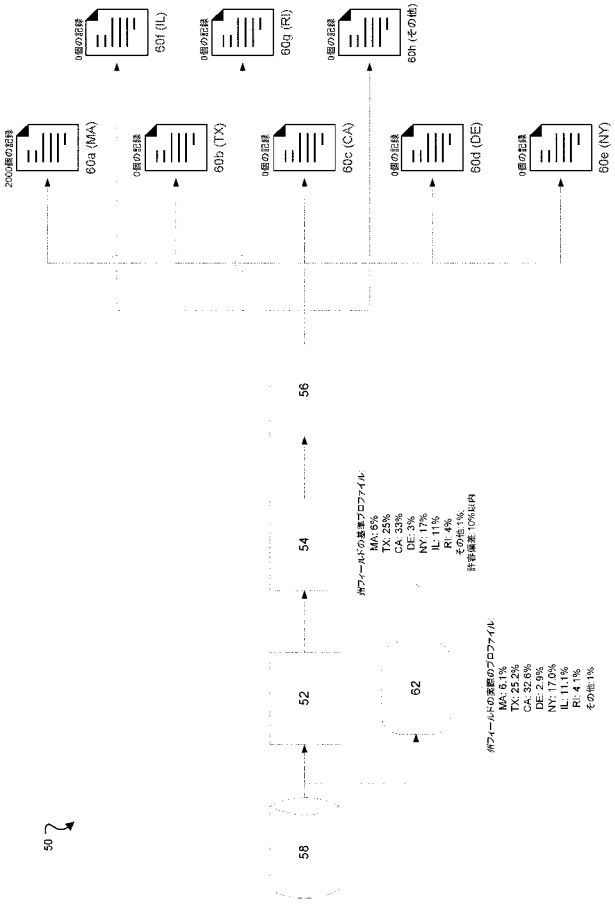
【図８Ｃ】

顧客_ID	州	購入_価格
12575	CA	127.51
35468	MA	547.77
26842	NY	29.80
32458	TX	245.01
54893	IL	55.76
05786	TX	18.76
65420	CA	442.10
12475	CA	54.30
54978	NY	22.99
97845	IR	65.17
65432	IL	18.19
54983	CA	25.44
32877	TX	509.60
54446	TX	63.12
23158	CA	45.18
35458	NY	56.04
87653	IR	17.10
65487	NY	152.60
32455	CA	146.12
54877	IL	33.32
78361	TX	423.01
12756	DE	18.76
45886	TX	87.45
32115	CA	108.69
68575	CA	146.08

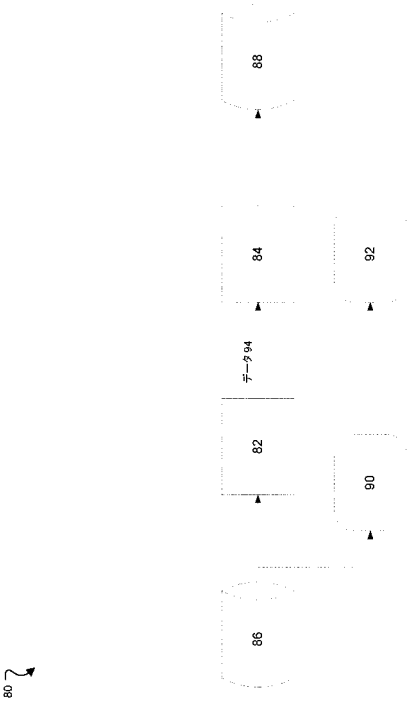
【図９Ａ】



【図９Ｂ】



【図１０Ａ】

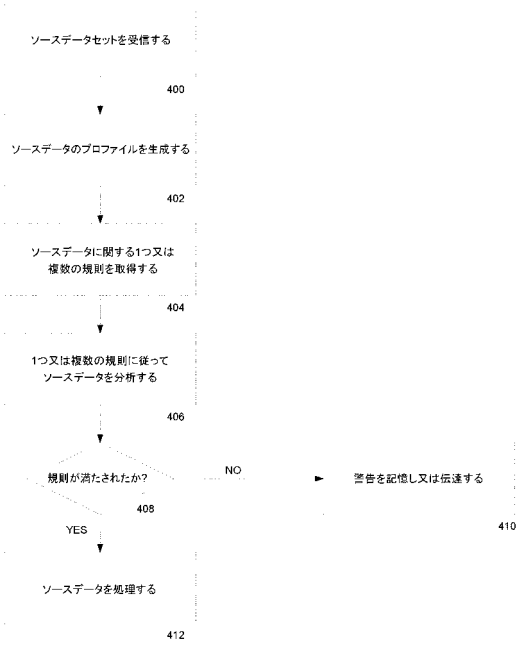


【図１０Ｂ】

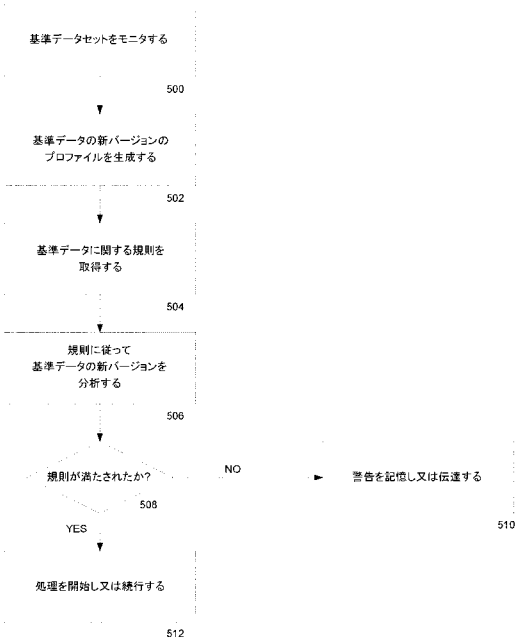
94

電話_番号	開始_時間	持続時間
6508001245	12:45:00 PM	0:30
4082604525	12:45:08 PM	1:15
212508412E	12:45:12 PM	10:12
6174438771	12:45:18 PM	8:07
3024023548	12:45:22 PM	0:02
21445893546	12:45:25 PM	12:17
9724505465	12:45:30 PM	8:32
917508459E	12:45:30 PM	0:09
7738645947	12:45:32 PM	1:12
2105782156	12:45:35 PM	4:01
4085647895	12:45:38 PM	0:14
6174546824	12:45:40 PM	2:08
214568778F	12:45:41 PM	14:16
7735684456	12:45:42 PM	9:12
6175443215	12:45:45 PM	58:17
212565788E	12:45:46 PM	7:26
3025488865	12:45:49 PM	0:16
9175453256	12:45:50 PM	1:38

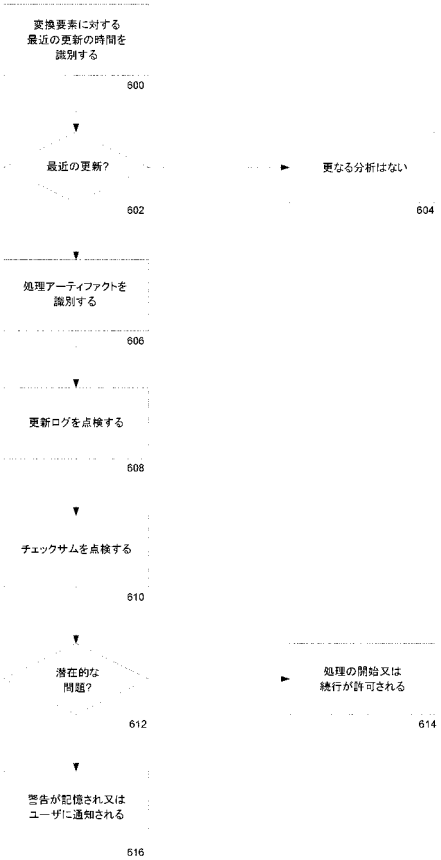
【図１１】



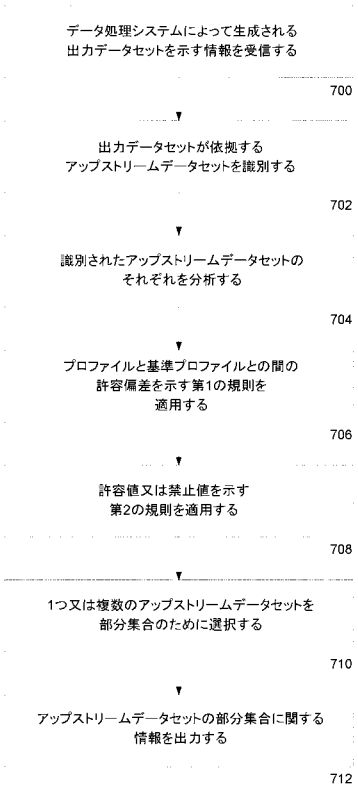
【図１２】



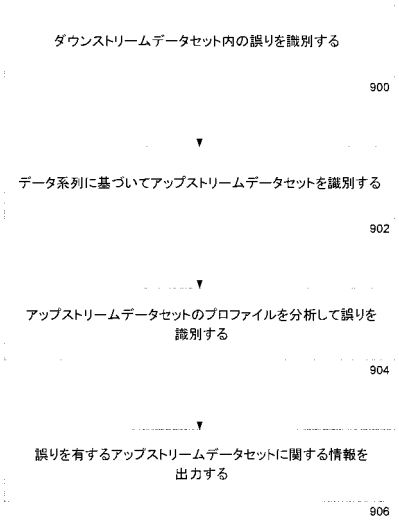
【図１３】



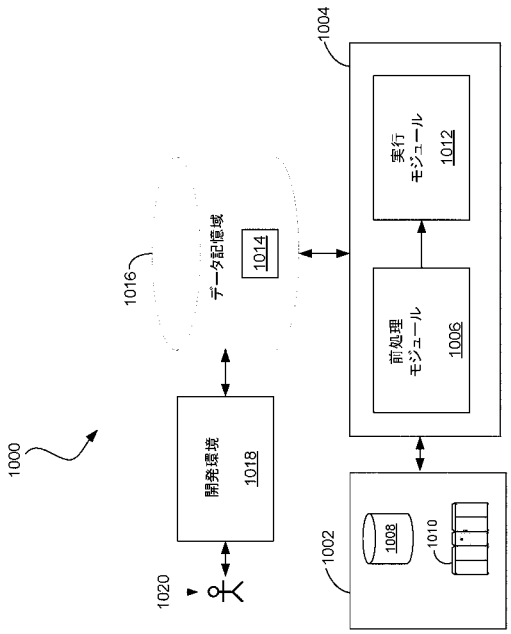
【図１４】


データ処理システムによって生成される
出力データセットを示す情報を受信する
700
出力データセットが依拠する
アップストリームデータセットを識別する
702
識別されたアップストリームデータセットの
それぞれを分析する
704
プロファイルと基準プロファイルとの間の
許容偏差を示す第1の規則を
適用する
706
許容値又は禁止値を示す
第2の規則を適用する
708
1つ又は複数のアップストリームデータセットを
部分集合のために選択する
710
アップストリームデータセットの部分集合に関する
情報を出力する
712

【図１５】


ダウンストリームデータセット内の誤りを識別する
900
データ系列に基づいてアップストリームデータセットを識別する
902
アップストリームデータセットのプロファイルを分析して誤りを
識別する
904
誤りを有するアップストリームデータセットに関する情報を
出力する
906

【図１６】


1000
1002
1004
1006
前処理
モジュール
1008
1010
1012
実行
モジュール
1014
データ記憶域
1016
1018
開発環境
1020

【手続補正書】
【提出日】令和2年6月12日(2020.6.12)
【手続補正１】
【補正対象書類名】特許請求の範囲
【補正対象項目名】全文
【補正方法】変更
【補正の内容】
【特許請求の範囲】
【請求項１】
データセットのフィールドについてのデータ品質規則を決定するためのコンピュータ実装方法であって、
前記データセットを分析することであって、前記データセットのフィールドについての基準プロファイルを決定することを含む、ことと、
決定された前記基準プロファイルに基づいて、前記データセットの前記フィールドについてのデータ品質規則を決定することであって、前記データセットの前記フィールドについてのデータ品質規則は、
（ｉ）前記フィールドについての前記基準プロファイルと、前記データセットの一つ以上のデータ記録の前記フィールドについてのプロファイルとの間の許容偏差
（ｉｉ）前記データセットのデータ記録の前記フィールドのデータ要素についての許容値、又は
（ｉｉｉ）前記データセットのデータ記録の前記フィールドのデータ要素についての禁止値の一つ以上を示す、ことと、
を含むコンピュータ実装方法。
【請求項２】
請求項１に記載の方法であって、前記データセットを分析することは、前記データセットの一つ以上の履歴インスタンスを分析することを含む、方法。
【請求項３】
請求項１に記載の方法であって、前記フィールドについての前記基準プロファイルを決定することは、前記フィールドについての履歴平均プロファイルを決定することを含む、方法。
【請求項４】
請求項３に記載の方法であって、前記データセットを分析することは、前記フィールドのデータ要素についての前記履歴平均プロファイルの変動が閾値量より少なく変動するまで前記データセットの複数の特定のインスタンスを分析することを含む、方法。
【請求項５】
請求項１に記載の方法であって、前記フィールドについての前記基準プロファイルを決定することは、前記フィールドのデータ要素についての履歴平均値を識別することを含む、方法。
【請求項６】
請求項１に記載の方法であって、前記フィールドについての前記基準プロファイルを決定することは、前記フィールドのデータ要素についての標準偏差値を識別することを含む、方法。
【請求項７】
請求項１に記載の方法であって、前記フィールドについての前記基準プロファイルを決定することは、前記フィールドのデータ要素についての個別値の数を識別することを含む、方法。
【請求項８】
請求項１に記載の方法であって、前記データセットを分析することは、前記データセットの予め定められた数の特定のインスタンスを分析することを含む、方法。
【請求項９】

請求項１に記載の方法であって、機械学習技術を用いて、前記データセットを分析することを含む、方法。
【請求項１０】
請求項１に記載の方法であって、前記データ品質規則を前記データセットの特定のインスタンスの前記一つ以上のデータ記録に適用することを含む、方法。
【請求項１１】
請求項１０に記載の方法であって、前記データ品質規則を前記データセットの前記特定のインスタンスに適用することは、前記データセットの前記特定のインスタンスが誤り又は起こり得る誤りを有するものと判定することを含む、方法。
【請求項１２】
請求項１１に記載の方法であって、前記データセットの前記特定のインスタンスが誤り又は起こり得る誤りを有するものと判定することは、
前記フィールドについての前記基準プロファイルと、前記データセットの特定のインスタンスの一つ以上のデータ記録の前記フィールドについてのプロファイルとの間の偏差を判定することと、
前記基準プロファイルと前記プロファイルとの間の前記偏差が前記許容偏差を超えるものと判定することと
を含む、方法。
【請求項１３】
請求項１１に記載の方法であって、前記データセットの前記特定のインスタンスが誤り又は起こり得る誤りを有するものと判定することは、前記許容値又は前記禁止を満たさない前記データセットの前記特定のインスタンスの一つ以上のデータ記録のフィールドのデータ要素を識別することを含む、方法。
【請求項１４】
データセットのフィールドについてのデータ品質規則を決定することを計算システムに行わせる命令を記憶する非一時的コンピュータ可読媒体であって、前記命令は、前記計算システムに、
前記データセットを分析することであって、前記データセットのフィールドについての基準プロファイルを決定することを含む、ことと、
決定された前記基準プロファイルに基づいて、前記データセットの前記フィールドについてのデータ品質規則を決定することであって、前記データセットの前記フィールドについてのデータ品質規則は、
（ｉ）前記フィールドについての前記基準プロファイルと、前記データセットの一つ以上のデータ記録の前記フィールドについてのプロファイルとの間の許容偏差
（ｉｉ）前記データセットのデータ記録の前記フィールドのデータ要素についての許容値、又は
（ｉｉｉ）前記データセットのデータ記録の前記フィールドのデータ要素についての禁止値の一つ以上を示す、ことと、
を行わせる、非一時的コンピュータ可読媒体。
【請求項１５】
請求項１４に記載の非一時的コンピュータ可読媒体であって、前記データセットを分析することは、前記データセットの一つ以上の履歴インスタンスを分析することを含む、非一時的コンピュータ可読媒体。
【請求項１６】
請求項１４に記載の非一時的コンピュータ可読媒体であって、前記フィールドについての前記基準プロファイルを決定することは、前記フィールドについての履歴平均プロファイルを決定することを含む、非一時的コンピュータ可読媒体。
【請求項１７】
請求項１６に記載の非一時的コンピュータ可読媒体であって、前記データセットを分析することは、前記フィールドのデータ要素についての前記履歴平均プロファイルの変動が

閾値量より少なく変動するまで前記データセットの複数の特定のインスタンスを分析することを含む、非一時的コンピュータ可読媒体。
【請求項１８】
請求項１４に記載の非一時的コンピュータ可読媒体であって、前記フィールドについての前記基準プロファイルを決定することは、前記フィールドのデータ要素についての履歴平均値を識別することを含む、非一時的コンピュータ可読媒体。
【請求項１９】
請求項１４に記載の非一時的コンピュータ可読媒体であって、前記フィールドについての前記基準プロファイルを決定することは、前記フィールドのデータ要素についての標準偏差値を識別することを含む、非一時的コンピュータ可読媒体。
【請求項２０】
請求項１４に記載の非一時的コンピュータ可読媒体であって、前記フィールドについての前記基準プロファイルを決定することは、前記フィールドのデータ要素についての個別値の数を識別することを含む、非一時的コンピュータ可読媒体。
【請求項２１】
請求項１４に記載の非一時的コンピュータ可読媒体であって、前記データセットを分析することは、前記データセットの予め定められた数の特定のインスタンスを分析することを含む、非一時的コンピュータ可読媒体。
【請求項２２】
請求項１４に記載の非一時的コンピュータ可読媒体であって、前記命令は、前記計算システムに、機械学習技術を用いて、前記データセットを分析させる、非一時的コンピュータ可読媒体。
【請求項２３】
請求項１４に記載の非一時的コンピュータ可読媒体であって、前記命令は、前記計算システムに、前記データ品質規則を前記データセットの特定のインスタンスの前記一つ以上のデータ記録に適用させる、非一時的コンピュータ可読媒体。
【請求項２４】
請求項２３に記載の非一時的コンピュータ可読媒体であって、前記データ品質規則を前記データセットの前記特定のインスタンスに適用することは、前記データセットの前記特定のインスタンスが誤り又は起こり得る誤りを有するものと判定することを含む、非一時的コンピュータ可読媒体。
【請求項２５】
請求項２４に記載の非一時的コンピュータ可読媒体であって、前記データセットの前記特定のインスタンスが誤り又は起こり得る誤りを有するものと判定することは、
前記フィールドについての前記基準プロファイルと、前記データセットの特定のインスタンスの一つ以上のデータ記録の前記フィールドについてのプロファイルとの間の偏差を判定することと、
前記基準プロファイルと前記プロファイルとの間の前記偏差が前記許容偏差を超えるものと判定することと
を含む、非一時的コンピュータ可読媒体。
【請求項２６】
請求項２４に記載の非一時的コンピュータ可読媒体であって、前記データセットの前記特定のインスタンスが誤り又は起こり得る誤りを有するものと判定することは、前記許容値又は前記禁止を満たさない前記データセットの前記特定のインスタンスの一つ以上のデータ記録のフィールドのデータ要素を識別することを含む、非一時的コンピュータ可読媒体。
【請求項２７】
データセットのフィールドについてのデータ品質規則を決定するための計算システムであって、
メモリに結合されている一つ以上のプロセッサであって、前記一つ以上のプロセッサ及

び前記メモリは、
前記データセットを分析することであって、前記データセットのフィールドについての基準プロファイルを決定することを含む、ことと、
決定された前記基準プロファイルに基づいて、前記データセットの前記フィールドについてのデータ品質規則を決定することであって、前記データセットの前記フィールドについてのデータ品質規則は、
（ｉ）前記フィールドについての前記基準プロファイルと、前記データセットの一つ以上のデータ記録の前記フィールドについてのプロファイルとの間の許容偏差
（ｉｉ）前記データセットのデータ記録の前記フィールドのデータ要素についての許容値、又は
（ｉｉｉ）前記データセットのデータ記録の前記フィールドのデータ要素についての禁止値の一つ以上を示す、ことと、
を行うように構成されている、計算システム。
【請求項２８】
請求項２７に記載の計算システムであって、前記データセットを分析することは、前記データセットの一つ以上の履歴インスタンスを分析することを含む、計算システム。
【請求項２９】
請求項２７に記載の計算システムであって、前記フィールドについての前記基準プロファイルを決定することは、前記フィールドについての履歴平均プロファイルを決定することを含む、計算システム。
【請求項３０】
請求項２９に記載の計算システムであって、前記データセットを分析することは、前記フィールドのデータ要素についての前記履歴平均プロファイルの変動が閾値量より少なく変動するまで前記データセットの複数の特定のインスタンスを分析することを含む、計算システム。
【請求項３１】
請求項２７に記載の計算システムであって、前記フィールドについての前記基準プロファイルを決定することは、前記フィールドのデータ要素についての履歴平均値を識別することを含む、計算システム。
【請求項３２】
請求項２７に記載の計算システムであって、前記フィールドについての前記基準プロファイルを決定することは、前記フィールドのデータ要素についての標準偏差値を識別することを含む、計算システム。
【請求項３３】
請求項２７に記載の計算システムであって、前記フィールドについての前記基準プロファイルを決定することは、前記フィールドのデータ要素についての個別値の数を識別することを含む、計算システム。
【請求項３４】
請求項２７に記載の計算システムであって、前記データセットを分析することは、前記データセットの予め定められた数の特定のインスタンスを分析することを含む、計算システム。
【請求項３５】
請求項２７に記載の計算システムであって、前記一つ以上のプロセッサ及び前記メモリは、機械学習技術を用いて、前記データセットを分析するように構成されている、計算システム。
【請求項３６】
請求項２７に記載の計算システムであって、前記一つ以上のプロセッサ及び前記メモリは、前記データ品質規則を前記データセットの特定のインスタンスの前記一つ以上のデータ記録に適用するように構成されている、計算システム。
【請求項３７】

請求項３６に記載の計算システムであって、前記データ品質規則を前記データセットの前記特定のインスタンスに適用することは、前記データセットの前記特定のインスタンスが誤り又は起こり得る誤りを有するものと判定することを含む、計算システム。

【請求項３８】

請求項３７に記載の計算システムであって、前記データセットの前記特定のインスタンスが誤り又は起こり得る誤りを有するものと判定することは、

前記フィールドについての前記基準プロファイルと、前記データセットの特定のインスタンスの一つ以上のデータ記録の前記フィールドについてのプロファイルとの間の偏差を判定することと、

前記基準プロファイルと前記プロファイルとの間の前記偏差が前記許容偏差を超えるものと判定することと

を含む、計算システム。

【請求項３９】

請求項３７に記載の計算システムであって、前記データセットの前記特定のインスタンスが誤り又は起こり得る誤りを有するものと判定することは、前記許容値又は前記禁止を満たさない前記データセットの前記特定のインスタンスの一つ以上のデータ記録のフィールドのデータ要素を識別することを含む、計算システム。

【請求項４０】

データセットのフィールドについてのデータ品質規則を決定するための計算システムであって、

前記データセットを分析する手段であって、前記データセットのフィールドについての基準プロファイルを決定することを含む、手段と、

決定された前記基準プロファイルに基づいて、前記データセットの前記フィールドについてのデータ品質規則を決定する手段であって、前記データセットの前記フィールドについてのデータ品質規則は、

（ｉ）前記フィールドについての前記基準プロファイルと、前記データセットの一つ以上のデータ記録の前記フィールドについてのプロファイルとの間の許容偏差

（ｉｉ）前記データセットのデータ記録の前記フィールドのデータ要素についての許容値、又は

（ｉｉｉ）前記データセットのデータ記録の前記フィールドのデータ要素についての禁止値の一つ以上を示す、手段と、

を備える、計算システム。

フロントページの続き

１．ＵＮＩＸ
２．ＪＡＶＡ

(72)発明者　スピッツ，チャック
アメリカ合衆国，マサチューセッツ州　０２４８２，ウェルズリー，アイビー　ロード　３３
(72)発明者　ゴウルド，ジョエル
アメリカ合衆国，マサチューセッツ州　０２４７４，アーリントン，リー　テラス　２７

Ｆターム(参考) 5B175 FB04　KA09

【外国語明細書】
2020161147000001.pdf