(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第5690349号
(P5690349)
(45)発行日　平成27年3月25日(2015.3.25)　(24)登録日　平成27年2月6日(2015.2.6)

(51)Int.Cl.　F I
G06F 12/00 (2006.01)　G06F 12/00 511C

請求項の数 42 (全 21 頁)

(21)出願番号　特願2012-539018(P2012-539018)
(86)(22)出願日　平成22年11月12日(2010.11.12)
(65)公表番号　特表2013-511097(P2013-511097A)
(43)公表日　平成25年3月28日(2013.3.28)
(86)国際出願番号　PCT/US2010/056530
(87)国際公開番号　WO2011/060257
(87)国際公開日　平成23年5月19日(2011.5.19)
審査請求日　平成25年11月11日(2013.11.11)
(31)優先権主張番号　61/260,997
(32)優先日　平成21年11月13日(2009.11.13)
(33)優先権主張国　米国(US)

(73)特許権者　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
(72)発明者　パーメンター，デイビッド，ダブリュー．
アメリカ合衆国，マサチューセッツ州　０２４５８，ニュートン，ヒュンウェル　アベニュー　１６５

最終頁に続く

(54)【発明の名称】レコード形式情報の管理

(57)【特許請求の範囲】
【請求項１】
データ記憶システム内の形式情報に基づいてデータ処理システム内で処理するためのデータを作成する方法であって、
入力デバイス又はポートを介してそれぞれのフィールドに対する１つ又は複数の値をそれぞれが有するレコードを含むデータを受信するステップと、
前記データ処理システム内の前記データを処理するためのターゲットレコード形式を決定するステップであって、
前記データ内の複数のレコードそれぞれを、前記データが前記データ記憶システム内に格納された１つ又は複数の候補レコード形式と一致するか否かを決定するために、複数の妥当性確認テストに従って分析すること、各候補レコード形式は、１つ又は複数のフィールドのグループの各フィールドについて形式を指定し、各妥当性確認テストは、前記データ記憶システム内に格納された少なくとも１つの候補レコード形式に対応し、及び
前記妥当性確認テストの結果の受信に応答して、前記ターゲットレコード形式を、選択済み候補レコード形式に対応する少なくとも１つの妥当性確認テストに従って少なくとも部分的一致が決定された選択済み候補レコード形式と、前記データに関連付けられた既知のデータタイプに従って選択されたパーザによって生成された解析済みレコード形式と、前記データの特徴の分析から生成された構築済みレコード形式のうち少なくとも１つに基づいて前記データに関連付けること
を含むステップと

を含む、方法。
【請求項２】
前記ターゲットレコード形式は、いずれの前記妥当性確認テストも前記候補レコード形式のうちの１つ又は複数に対する少なくとも部分的一致を決定しないことに応答して、前記解析済みレコード形式に基づいて前記データに関連付けられる、請求項１に記載の方法。
【請求項３】
前記データに関連付けられた前記既知のデータタイプは、前記データのファイルタイプに基づいて知られる、請求項２に記載の方法。
【請求項４】
前記データの前記ファイルタイプは、ファイル拡張子に対応する、請求項３に記載の方法。
【請求項５】
前記ターゲットレコード形式は、いずれの前記妥当性確認テストも前記候補レコード形式のうちの１つ又は複数に対する少なくとも部分的一致を決定しないこと、及び前記データに関連付けられた既知のデータタイプを有さないことに応答して、前記構築済みレコード形式に基づいて前記データに関連付けられる、請求項１に記載の方法。
【請求項６】
前記データの特徴の分析から前記構築済みレコード形式を生成するステップは、前記データ内のタグを認識すること、及び前記認識されたタグに基づいて複数のレコードを決定するために前記データを解析することを含む、請求項５に記載の方法。
【請求項７】
前記データの特徴の分析から前記構築済みレコード形式を生成するステップは、前記データ内の区切り文字を認識すること、及び前記認識された区切り文字に基づいて複数のレコードを決定するために前記データを解析することを含む、請求項５に記載の方法。
【請求項８】
前記データの特徴の分析から前記構築済みレコード形式を生成するステップは、前記データが、複数のレコードの値を示すタグ又は区切り文字のない実質上バイナリ形式であることを認識すること、及びユーザインターフェイスから１つ又は複数のフィールド識別子を受け取ることを含む、請求項５に記載の方法。
【請求項９】
第１の候補レコード形式に対応する前記複数の妥当性確認テストのうちの第１の妥当性確認テストに従って、前記データ内の前記複数のレコードを分析することは、各フィールドに対する前記第１の候補レコード形式によって指定された前記形式で各レコードに対する値を決定するために、前記第１の候補レコード形式を前記データに適用することを含む、請求項１に記載の方法。
【請求項１０】
前記データが前記第１の候補レコード形式と一致するか否かを決定するステップは、幾つかの有効値が所定のしきい値よりも大きいか否かを決定するために、前記第１の妥当性確認テストに従って前記複数のレコードに対して前記決定された値を分析することを含む、請求項９に記載の方法。
【請求項１１】
前記第１の妥当性確認テストに従って前記複数のレコードのうちの第１のレコードに対して決定された値を分析するステップは、各フィールドに対して決定された各値について対応するフィールドテストを実行することを含む、請求項１０に記載の方法。
【請求項１２】
第１のフィールドに対して決定された値について第１のフィールドテストを実行するステップは、前記決定された値の中の幾つかの文字を所定数の文字と突き合せることを含む、請求項１１に記載の方法。
【請求項１３】

第１のフィールドに対して決定された値について第１のフィールドテストを実行するステップは、前記決定された値を前記第１のフィールドに対する複数の所定の有効値のうちの１つと突き合せることを含む、請求項１１に記載の方法。

【請求項１４】

前記有効値の数は、所与のフィールドに対して決定された値が前記所与のフィールドに対応する前記フィールドテストに合格した幾つかのレコードに基づくものである、請求項１１に記載の方法。

【請求項１５】

データ記憶システム内の形式情報に基づいてデータ処理システム内で処理するためにデータを準備するためのシステムであって、

入力デバイス又はポートを介してそれぞれのフィールドに対する１つ又は複数の値をそれぞれが有するレコードを含むデータを受信するための手段と、

前記データ処理システム内の前記データを処理するためのターゲットレコード形式を決定するための手段と、を含み、

前記決定するための手段は、

前記データ内の複数のレコードそれぞれを、前記データが前記データ記憶システム内に格納された１つ又は複数の候補レコード形式と一致するか否かを決定するために、複数の妥当性確認テストに従って分析すること、各候補レコード形式は、１つ又は複数のフィールドのグループの各フィールドについて形式を指定し、各妥当性確認テストは、前記データ記憶システム内に格納された少なくとも１つの候補レコード形式に対応し、及び

前記妥当性確認テストの結果の受信に応答して、前記ターゲットレコード形式を、選択済み候補レコード形式に対応する少なくとも１つの妥当性確認テストに従って少なくとも部分的一致が決定された選択済み候補レコード形式と、前記データに関連付けられた既知のデータタイプに従って選択されたパーザによって生成された解析済みレコード形式と、前記データの特徴の分析から生成された構築済みレコード形式のうち少なくとも１つに基づいてデータに関連付けること、を含む、

システム。

【請求項１６】

データ記憶システム内の形式情報に基づいてデータ処理システム内で処理するためにデータを準備するためのコンピュータプログラムを格納するコンピュータ読み取り可能媒体であって、前記コンピュータプログラムが、

入力デバイス又はポートを介してそれぞれのフィールドに対する１つ又は複数の値をそれぞれが有するレコードを含むデータを受信するステップと、

前記データ処理システム内の前記データを処理するためのターゲットレコード形式を決定するステップと、をコンピュータに実行させるための命令を含み、

前記決定するステップは、

前記データ内の複数のレコードそれぞれを、前記データが前記データ記憶システム内に格納された１つ又は複数の候補レコード形式と一致するか否かを決定するために、複数の妥当性確認テストに従って分析すること、各候補レコード形式は、１つ又は複数のフィールドのグループの各フィールドについて形式を指定し、各妥当性確認テストは、前記データ記憶システム内に格納された少なくとも１つの候補レコード形式に対応し、及び

前記妥当性確認テストの結果の受信に応答して、前記ターゲットレコード形式を、選択済み候補レコード形式に対応する少なくとも１つの妥当性確認テストに従って少なくとも部分的一致が決定された選択済み候補レコード形式と、前記データに関連付けられた既知のデータタイプに従って選択されたパーザによって生成された解析済みレコード形式と、前記データの特徴の分析から生成された構築済みレコード形式のうち少なくとも１つに基づいてデータに関連付けること、を含む、

コンピュータ読み取り可能媒体。

【請求項１７】

前記ターゲットレコード形式は、いずれの前記妥当性確認テストも前記候補レコード形

式のうちの１つ又は複数に対する少なくとも部分的一致を決定しないことに応答して、前記解析済みレコード形式に基づいて前記データに関連付けられる、請求項１５に記載のシステム。

【請求項１８】

前記データに関連付けられた前記既知のデータタイプは、前記データのファイルタイプに基づいて知られる、請求項１７に記載のシステム。

【請求項１９】

前記データの前記ファイルタイプは、ファイル拡張子に対応する、請求項１８に記載のシステム。

【請求項２０】

前記ターゲットレコード形式は、いずれの前記妥当性確認テストも前記候補レコード形式のうちの１つ又は複数に対する少なくとも部分的一致を決定しないこと、及び前記データに関連付けられた既知のデータタイプを有さないことに応答して、前記構築済みレコード形式に基づいて前記データに関連付けられる、請求項１５に記載のシステム。

【請求項２１】

前記データの特徴の分析から前記構築済みレコード形式を生成することは、前記データ内のタグを認識すること、及び前記認識されたタグに基づいて複数のレコードを決定するために前記データを解析することを含む、請求項２０に記載のシステム。

【請求項２２】

前記データの特徴の分析から前記構築済みレコード形式を生成することは、前記データ内の区切り文字を認識すること、及び前記認識された区切り文字に基づいて複数のレコードを決定するために前記データを解析することを含む、請求項２０に記載のシステム。

【請求項２３】

前記データの特徴の分析から前記構築済みレコード形式を生成することは、前記データが、複数のレコードの値を示すタグ又は区切り文字のない実質上バイナリ形式であることを認識すること、及びユーザインターフェイスから１つ又は複数のフィールド識別子を受け取ることを含む、請求項２０に記載のシステム。

【請求項２４】

第１の候補レコード形式に対応する前記複数の妥当性確認テストのうちの第１の妥当性確認テストに従って、前記データ内の前記複数のレコードを分析することは、各フィールドに対する前記第１の候補レコード形式によって指定された前記形式で各レコードに対する値を決定するために、前記第１の候補レコード形式を前記データに適用することを含む、請求項１５に記載のシステム。

【請求項２５】

前記データが前記第１の候補レコード形式と一致するか否かを決定することは、幾つかの有効値が所定のしきい値よりも大きいか否かを決定するために、前記第１の妥当性確認テストに従って前記複数のレコードに対して前記決定された値を分析することを含む、請求項２４に記載のシステム。

【請求項２６】

前記第１の妥当性確認テストに従って前記複数のレコードのうちの第１のレコードに対して決定された値を分析することは、各フィールドに対して決定された各値について対応するフィールドテストを実行することを含む、請求項２５に記載のシステム。

【請求項２７】

第１のフィールドに対して決定された値について第１のフィールドテストを実行することは、前記決定された値の中の幾つかの文字を所定数の文字と突き合せることを含む、請求項２６に記載のシステム。

【請求項２８】

第１のフィールドに対して決定された値について第１のフィールドテストを実行することは、前記決定された値を前記第１のフィールドに対する複数の所定の有効値のうちの１つと突き合せることを含む、請求項２６に記載のシステム。

【請求項２９】
前記有効値の数は、所与のフィールドに対して決定された値が前記所与のフィールドに対応する前記フィールドテストに合格した幾つかのレコードに基づくものである、請求項２６に記載のシステム。
【請求項３０】
前記ターゲットレコード形式は、いずれの前記妥当性確認テストも前記候補レコード形式のうちの１つ又は複数に対する少なくとも部分的一致を決定しないことに応答して、前記解析済みレコード形式に基づいて前記データに関連付けられる、請求項１６に記載の媒体。
【請求項３１】
前記データに関連付けられた前記既知のデータタイプは、前記データのファイルタイプに基づいて知られる、請求項３０に記載の媒体。
【請求項３２】
前記データの前記ファイルタイプは、ファイル拡張子に対応する、請求項３１に記載の媒体。
【請求項３３】
前記ターゲットレコード形式は、いずれの前記妥当性確認テストも前記候補レコード形式のうちの１つ又は複数に対する少なくとも部分的一致を決定しないこと、及び前記データに関連付けられた既知のデータタイプを有さないことに応答して、前記構築済みレコード形式に基づいて前記データに関連付けられる、請求項１６に記載の媒体。
【請求項３４】
前記データの特徴の分析から前記構築済みレコード形式を生成するステップは、前記データ内のタグを認識すること、及び前記認識されたタグに基づいて複数のレコードを決定するために前記データを解析することを含む、請求項３３に記載の媒体。
【請求項３５】
前記データの特徴の分析から前記構築済みレコード形式を生成するステップは、前記データ内の区切り文字を認識すること、及び前記認識された区切り文字に基づいて複数のレコードを決定するために前記データを解析することを含む、請求項３３に記載の媒体。
【請求項３６】
前記データの特徴の分析から前記構築済みレコード形式を生成するステップは、前記データが、複数のレコードの値を示すタグ又は区切り文字のない実質上バイナリ形式であることを認識すること、及びユーザインターフェイスから１つ又は複数のフィールド識別子を受け取ることを含む、請求項３３に記載の媒体。
【請求項３７】
第１の候補レコード形式に対応する前記複数の妥当性確認テストのうちの第１の妥当性確認テストに従って、前記データ内の前記複数のレコードを分析することは、各フィールドに対する前記第１の候補レコード形式によって指定された前記形式で各レコードに対する値を決定するために、前記第１の候補レコード形式を前記データに適用することを含む、請求項１６に記載の媒体。
【請求項３８】
前記データが前記第１の候補レコード形式と一致するか否かを決定するステップは、幾つかの有効値が所定のしきい値よりも大きいか否かを決定するために、前記第１の妥当性確認テストに従って前記複数のレコードに対して前記決定された値を分析することを含む、請求項３７に記載の媒体。
【請求項３９】
前記第１の妥当性確認テストに従って前記複数のレコードのうちの第１のレコードに対して決定された値を分析するステップは、各フィールドに対して決定された各値について対応するフィールドテストを実行することを含む、請求項３８に記載の媒体。
【請求項４０】
第１のフィールドに対して決定された値について第１のフィールドテストを実行するス

テップは、前記決定された値の中の幾つかの文字を所定数の文字と突き合せることを含む、請求項３９に記載の媒体。
【請求項４１】
第１のフィールドに対して決定された値について第１のフィールドテストを実行するステップは、前記決定された値を前記第１のフィールドに対する複数の所定の有効値のうちの１つと突き合せることを含む、請求項３９に記載の媒体。
【請求項４２】
前記有効値の数は、所与のフィールドに対して決定された値が前記所与のフィールドに対応する前記フィールドテストに合格した幾つかのレコードに基づくものである、請求項３９に記載の媒体。
【発明の詳細な説明】
【技術分野】
【０００１】
（関連出願への相互参照）
本出願は、２００９年１１月１３日出願の米国特許出願第６１／２６０，９９７号に対する優先権を主張し、参照によりその内容を本明細書に組み込むものとする。
【０００２】
本明細書は、レコード形式情報の管理に関する。
【背景技術】
【０００３】
組織は、複数の異なるシステムからのデータを管理する。システムは、データのデータセットをシステムに固有の形式で生成してもよい。他のシステムは、カンマ区切りファイル又はＸＭＬ文書などの標準の形式を使用してデータセットを生成する。一般に、たとえデータセットの形式が標準的であっても、データセット内のレコード及びフィールドはシステム特有である。
【発明の概要】
【課題を解決するための手段】
【０００４】
幾つかのシステムは、インポート機構を通して他のシステムによって提供されたデータセットを受け入れる。インポートは、処理のために、外部データセットをシステム固有の形式に変換する。他のシステムは、システムが必ずしも変換を必要とせずに外部データセットを処理できるようにするのに十分なデータセットを記述するレコード形式を作成する。
【０００５】
一態様では、一般に、データ記憶システム内の形式情報に基づいてデータ処理システム内で処理するためのデータを作成する方法である。入力デバイス又はポートを介してそれぞれのフィールドに対する１つ又は複数の値をそれぞれが有するレコードを含むデータが受信される。データ処理システム内のデータを処理するためのターゲットレコード形式が決定される。データ内の複数のレコードは、データがデータ記憶システム内に格納された１つ又は複数の候補レコード形式と一致するか否かを決定するために、複数の妥当性確認テストに従って分析される。各候補レコード形式は、１つ又は複数のフィールドのグループの各フィールドについて形式を指定し、各妥当性確認テストは、データ記憶システム内に格納された少なくとも１つの候補レコード形式に対応する。妥当性確認テストの結果の受信に応答して、ターゲットレコード形式は、選択済み候補レコード形式に対応する少なくとも１つの妥当性確認テストに従って少なくとも部分的一致が決定された選択済み候補レコード形式と、データに関連付けられた既知のデータタイプに従って選択されたパーザ（parser）によって生成された解析済みレコード形式と、データの特徴の分析から生成された構築済みレコード形式のうち少なくとも１つに基づいてデータに関連付けられる。
【０００６】
態様は、以下の機能のうち１つ又はそれ以上を含むことができる。

【０００７】
いずれの妥当性確認テストも候補レコード形式のうちの１つ又は複数に対する少なくとも部分的一致を決定しないことに応答して、解析済みレコード形式に基づいてターゲットレコード形式をデータに関連付ける。データに関連付けられた既知のデータタイプは、データのファイルタイプに基づいて知ることができる。データのファイルタイプは、ファイル拡張子に対応するものとすることができる。いずれの妥当性確認テストも候補レコード形式のうち１つ又は複数に対する少なくとも部分的一致を決定しないこと、及びデータに関連付けられた既知のデータタイプを有さないことに応答して、構築済みレコード形式に基づいてターゲットレコード形式をデータに関連付ける。データの特徴の分析から構築済みレコード形式を生成することは、データ内のタグを認識すること、及び認識されたタグに基づいて複数のレコードを決定するためにデータを解析することを含んでもよい。データの特徴の分析から構築済みレコード形式を生成することは、データ内の区切り文字を認識すること、及び認識された区切り文字に基づいて複数のレコードを決定するためにデータを解析することを含んでもよい。データの特徴の分析から構築済みレコード形式を生成することは、データが、複数のレコードの値を示すタグ又は区切り文字のない、実質上バイナリ形式であることを認識すること、及びユーザインターフェイスから１つ又は複数のフィールド識別子を受け取ることを含んでもよい。第１の候補レコード形式に対応する複数の妥当性確認テストのうちの第１の妥当性確認テストに従ってデータ内の複数のレコードを分析することは、各フィールドに対する第１の候補レコード形式によって指定された形式で各レコードに対する値を決定するために、第１の候補レコード形式をデータに適用することを含んでもよい。データが第１の候補レコード形式と一致するか否かを決定することは、幾つかの有効値が所定のしきい値よりも大きいか否かを決定するために、第１の妥当性確認テストに従って複数のレコードに対して決定された値を分析することを含んでもよい。第１の妥当性確認テストに従って複数のレコードのうちの第１のレコードに対して決定された値を分析することは、各フィールドに対して決定された各値について対応するフィールドテストを実行することを含んでもよい。第１のフィールドに対して決定された値について第１のフィールドテストを実行することは、決定された値の中の幾つかの文字を所定数の文字と突き合せることを含んでもよい。第１のフィールドに対して決定された値について第１のフィールドテストを実行することは、決定された値を第１のフィールドに対する複数の所定の有効値のうちの１つと突き合せることを含んでもよい。有効値の数は、所与のフィールドに対して決定された値が所与のフィールドに対応するフィールドテストに合格した幾つかのレコードに基づくものとすることができる。

【０００８】
別の態様では、一般に、データ記憶システム内の形式情報に基づいてデータ処理システム内で処理するためにデータを準備するためのシステムは、入力デバイス又はポートを介してそれぞれのフィールドに対する１つ又は複数の値をそれぞれが有するレコードを含むデータを受信するための手段と、データ処理システム内のデータを処理するためのターゲットレコード形式を決定するための手段とを含む。ターゲットレコード形式を決定するための手段は、データが、データ記憶システムに格納された１つ又は複数の候補レコード形式と一致するか否かを決定するために、複数の妥当性確認テストに従ってデータ内の複数のレコードを分析し、各候補レコード形式は１つ又は複数のフィールドのグループのうちの各フィールドに対して形式を指定し、各妥当性確認テストはデータ記憶システムに格納された少なくとも１つの候補レコード形式に対応し、並びに、妥当性確認テストの結果の受信に応答して、選択済み候補レコード形式に対応する少なくとも１つの妥当性確認テストに従って少なくとも部分的一致が決定された選択済み候補レコード形式と、データに関連付けられた既知のデータタイプに従って選択されたパーザによって生成された解析済みレコード形式と、データの特徴の分析から生成された構築済みレコード形式のうち少なくとも１つに基づいてターゲットレコード形式をデータに関連付けるように構成してもよい。

【０００９】

別の態様では、一般に、コンピュータ読み取り可能媒体が、データ記憶システム内の形式情報に基づいてデータ処理システム内で処理するためにデータを準備するためのコンピュータプログラムを格納する。コンピュータプログラムは、入力デバイス又はポートを介してそれぞれのフィールドに対する１つ又は複数の値をそれぞれが有するレコードを含むデータを受信すること、並びに、データが、データ記憶システムに格納された１つ又は複数の候補レコード形式と一致するか否かを決定するために、複数の妥当性確認テストに従ってデータ内の複数のレコードを分析し、各候補レコード形式は１つ又は複数のフィールドのグループのうちの各フィールドに対して形式を指定し、各妥当性確認テストはデータ記憶システムに格納された少なくとも１つの候補レコード形式に対応し、及び妥当性確認テストの結果の受信に応答して、選択済み候補レコード形式に対応する少なくとも１つの妥当性確認テストに従って、少なくとも部分的一致が決定された選択済み候補レコード形式と、データに関連付けられた既知のデータタイプに従って選択されたパーザによって生成された解析済みレコード形式と、データの特徴の分析から生成された構築済みレコード形式のうち少なくとも１つに基づいてターゲットレコード形式をデータに関連付けることを含むデータ処理システム内のデータを処理するためのターゲットレコード形式を決定することをコンピュータに実行させるための命令を含む。

【００１０】

本発明の他の特徴及び利点は、以下の説明及び特許請求の範囲から明らかとなろう。

【図面の簡単な説明】

【００１１】

【図１】グラフベースの計算を実行するためのシステムのブロック図である。

【図２】レコード形式情報を管理するための例示的手順のフローチャートである。

【図３】例示的事前処理モジュールのブロック図である。

【図４】サンプルデータに基づいてレコード形式を決定する事前処理モジュールの例示的処理を示すブロック図である。

【図５】サンプルデータに基づいてレコード形式の妥当性を確認する事前処理モジュールの例示的処理を示すブロック図である。

【図６】サンプルデータに基づいて既存のレコード形式を識別する事前処理モジュールの例示的処理を示すブロック図である。

【図７】パーザに基づいてレコード形式を生成する事前処理モジュールの例示的処理を示すブロック図である。

【図８Ａ】レコード形式情報を管理するための例示的手順のフローチャートである。

【図８Ｂ】レコード形式情報を管理するための例示的手順のフローチャートである。

【発明を実施するための形態】

【００１２】

図１は、レコード形式管理技術が使用できる例示的なデータ処理システム１００を示す。システム１００は、記憶装置などのデータの１つ又は複数のソース、又はオンラインデータストリームへの接続を含むことができるデータソース１０２を含み、そのそれぞれが、様々な記憶形式（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によって使用される固有の形式）のうちのいずれかでデータを格納することができる。実行環境１０４は、事前処理モジュール１０６及び実行モジュール１１２を含む。実行環境１０４は、ＵＮＩＸオペレーティングシステムなどの好適なオペレーティングシステムの制御の下で１つ又は複数の汎用コンピュータ上でホストされてもよい。例えば、実行環境１０４は、ローカル（例えばＳＭＰコンピュータなどのマルチプロセッサシステム）、又はローカル分散型（例えばクラスタ又はＭＰＰとして結合された複数のプロセッサ）、又はリモート、又はリモート分散型（例えば、ローカルエリアネットワーク（ＬＡＮ）及び／又はワイドエリアネットワーク（ＷＡＮ）を介して結合された複数のプロセッサ）、又はそれらの任意の組合せのいずれかの複数の中央処理ユニット（ＣＰＵ）を使用するコンピュータシステムの構成を含む多重ノード並列コンピューティング環境を含むことができる。幾つかの実施態様では、実行モジュール１

１２は、１つ又は複数のプロセッサ上で実行する並列オペレーティングシステムであってもよいオペレーティングシステムを提供し、事前処理モジュール１０６は、そのオペレーティングシステム内で実行するプログラムとして実行される。ユーザ１１５は、表示された出力を見ること、及びユーザインターフェイス内に入力することによって、実行環境１０８と対話する（interact）こともできる。

【００１３】

事前処理モジュール１０６は、それぞれのフィールドについてそれぞれ１つ又は複数の値を有するレコードを含むデータをデータソース１０２から受信し、実行モジュール１１２を使用してレコードを処理するためのターゲットレコード形式を決定する。例えば、事前処理モジュール１０６は、適切なターゲットレコード形式１１４がデータ記憶システム１１６内にすでに格納されていることを決定するか、又は格納されていない場合、ターゲットレコード形式１１４を生成して、生成されたターゲットレコード形式１１４をデータ記憶システム１１６内に格納する。データソース１０２及びデータ記憶システム１１６を提供する記憶装置は、例えば、実行環境１０４を実行中のコンピュータに接続された記憶媒体（例えばハードドライブ１０８）上に格納されるように、実行環境１０４に対してローカルとしてもよく、又は例えば、リモート接続を介して、実行環境１０４を実行中のコンピュータと通信しているリモートシステム（例えばメインフレーム１１０）上でホストされるように、実行環境１０４に対してリモートとしてもよい。

【００１４】

実行モジュール１１２は、決定されたターゲットレコード形式１１４を使用して、データソース１０２から受信したレコードを解釈及び処理する。また、レコードを処理するために、実行モジュール１１２によって実行されるプログラムを開発者１２０が開発可能な開発環境１１８もデータ記憶システム１１６にアクセス可能である。幾つかの実施態様では、開発環境１１８は、頂点（コンポーネント又はデータセット）間で有向リンク（作業要素の流れを表す）によって接続された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、このような環境は、参照により本明細書に組み込むものとする、「Managing Parameters for Graph-Based Applications」という名称の米国特許出願第２００７／００１１６６８号で、より詳細に説明されている。

【００１５】

事前処理モジュール１０６は、異なる形のデータベースシステムを含む様々なタイプ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、場合によってはヌル値を含むそれぞれのフィールドに対する値（「属性」又は「カラム」とも呼ばれる）を有するレコードとして編成されてもよい。データソースから最初にデータを読み取る場合、そのデータソースからのレコードのレコード構造を記述するターゲットレコード形式は知られていないが、幾つかの状況では、事前処理モジュール１０６は、そのデータソース内のレコードに関する幾つかの初期の形式情報から始めてもよい。事前処理モジュール１０６は、処理されるレコードが格納されたレコード形式によって記述されているか、又はレコード形式が生成されるかを決定するために、データ記憶システム１１６に格納されたレコード形式の集合を管理する。レコード形式は、別個の値を表すビット数、レコード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タイプ（例えば文字列、符号付き／符号なし整数）などの様々な特徴を含むことができる。

【００１６】

図２を参照すると、プロセス２２０に関するフローチャートは、レコード形式を管理するための事前処理モジュール１０６のうちの幾つかの動作（operation）を含む。機能（capabilities）の中でもとりわけ、事前処理モジュール１０６はデータを受け入れる２２２。データは、ファイル、データベース、ユーザインターフェイス、入力ポート、又は任意の他の入力デバイスを通して受信してもよい。他の情報の中でもとりわけ、事前処理モジュール１０６は、データソースからのレコードに関するレコード形式、又はデータソース１０２からの１つ又は複数のレコードを含むサンプルデータ、或いはその両方を受信してもよい。サンプルデータは、処理されることになるすべてのレコード又はレコードのサ

ブセットを含むことができる。事前処理モジュール１０６は、事前処理モジュール１０６がどの動作の実行を要求されたかの指示も受信してもよい。

【００１７】

事前処理モジュール１０６の動作は、プロセス経路を決定すること２２４も含む。事前処理モジュール１０６は、受信したサンプルデータのレコードを解釈するために、レコード形式を決定するための複数の方法を有してもよい。事前処理モジュール１０６は、サンプルデータに関する潜在的レコード形式が入力として提供されるか否かに基づいてどのプロセス経路が適切であるかを決定してもよい。システムの幾つかの実施態様では、事前処理モジュール１０６は、どのプロセス経路が好ましいかを示すデータを受け入れる。

【００１８】

１つのプロセス経路に沿って、事前処理モジュール１０６の動作は、以下でより詳細に説明するように、サンプルデータの分析に基づいてサンプルデータのターゲットレコード形式を決定すること２２６を含む。

【００１９】

別のプロセス経路に沿って、事前処理モジュール１０６の動作は、提供されたレコード形式と提供されたサンプルデータとの比較２２８に基づいて、サンプルデータのターゲットレコード形式を決定することを含む。幾つかのケースでは、事前処理モジュール１０６は、サンプルデータと、受け入れたサンプルデータに潜在的に対応する提供されたレコード形式（又は格納されたレコード形式に対する識別子）を受け入れる。事前処理モジュール１０６は、レコード形式がサンプルデータの構造を表すか否かを決定するために、提供又は識別されたレコード形式とサンプルデータとを比較する。

【００２０】

別のプロセス経路に沿って、事前処理モジュール１０６の動作は、提供されたサンプルデータに関するレコード形式を見つけること２３０に基づいて、サンプルデータのターゲットレコード形式を決定することを含む。幾つかのケースでは、事前処理モジュール１０６はサンプルデータを受け入れ、レコード形式のうちのいずれかがサンプルデータの構造を正しく表すか否かを発見するために、このデータと、レコード形式リポジトリ（例えばデータ記憶システム１１６内でホストされる）内の既存のレコード形式とを比較する。

【００２１】

動作は、１つ又は複数の潜在的ターゲットレコード形式をユーザに提示すること２３２も含む。１つ又は複数のレコード形式が決定されると、そのレコード形式をユーザに提示することができる。ユーザは、複数のレコード形式から単一のレコード形式を選択してもよい。ユーザは、レコード形式を修正してもよい。

【００２２】

動作は、ターゲットレコード形式の妥当性を確認すること２３４も含む。レコード形式が事前処理モジュール１０６によって受け入れられる前に、事前処理モジュールは、提供されたサンプルデータに照らしてレコード形式の妥当性を確認してもよい。

【００２３】

動作は、ターゲットレコード形式に対して調整を提案すること２３６も含む。提供されたサンプルデータをレコード形式が解析できない場合、事前処理モジュール１０６は、レコード形式とサンプルデータとの間の不整合を識別する。この不整合は、サンプルデータを解析する場合に発生したエラーを分析することによって識別してもよい。また、不整合は、サンプルデータ及びレコード形式を分析することによって識別してもよい。次に、プロセス２２０は、この不整合を修復するように提案する。幾つかの実施態様では、プロセス２２０は、サンプルデータに基づいてレコード形式を修正することを推奨してもよい。例えば、フィールドが整数表現（例えば、１、２、３、４などの整数値の２進表現）となることをレコード形式が予測し、サンプルデータ内のそのフィールドが形式化された日付の表現（例えば１／２１／２００８、２１／１／２００８、０１－ＪＡＮ－２００８など）を含む場合、プロセス２２０は調整を提案してもよい。整数フィールドは形式化された日付を保持できず、日付フィールドは整数を保持できないため、プロセス２２０は、フィ

ールドを、日付又は整数のいずれかを含むことができる文字列に修正することを提案してもよい。別の例では、プロセス２２０は、レコード形式によって受け入れられる有効値の範囲を拡大することを提案してもよい。

【００２４】

動作は、ターゲットレコード形式を格納すること２３８も含む。ターゲットレコード形式は、レコード形式リポジトリに格納してもよい。

【００２５】

図３を参照すると、データ処理システム内で処理するためにデータを準備するための事前処理モジュール３００は、データを受け入れるための機構を含む。幾つかのケースでは、入力データはデータベース３１０としてもよい。データベース３１０は、システム１００によって処理されることになるデータを含んでもよい。他のケースでは、データベース３１０は、システム１００によって処理されることになるデータのより大きなセットを表すデータのサンプルセットを含んでもよい。さらに他のケースでは、データベースはデータのレコード形式の記述を含んでもよい。他のケースでは、入力データはサンプルデータとレコード形式との組合せを含んでもよい。入力データは、リレーショナルデータベース、フラットファイル、又はポートを介して又は別の入力デバイスを介して受信されるデータなどの入力をレコード形式プロセス３０２内に提供するための別の機構を介してレコード形式プロセス３０２に通信してもよい。

【００２６】

レコード形式プロセス３０２は、入力データ３１０を受け入れ、ターゲットレコード形式を決定する。幾つかのケースでは、入力データは複数のレコードで構成されるサンプルデータを含み、各レコードは複数フィールドに対する値を含む。サンプルデータは、レコード形式を決定するために分析される。他のケースでは、サンプルデータは提供されたレコード形式と比較される。他のケースでは、サンプルデータは、最良の適合を決定するために、レコード形式リポジトリ内の既存のレコード形式と比較される。

【００２７】

幾つかのケースでは、レコード形式プロセス３０２は、ターゲットレコード形式を決定するために、いずれかの既存のパーザが入力データ３１０を解析できるか否かを決定するためのパーザを含むパーザカタログ３０６を検査する。入力データ３１０を処理するためのパーザが存在しない場合、レコード形式プロセス３０２は、ターゲットレコード形式を決定するための新しいパーザの構築を可能にするカスタムパーザビルダモジュール３０８にアクセスしてもよい。

【００２８】

ユーザには、レコード形式が提示され、レコード形式を調整するように許可してもよい。調整されたレコード形式は、レコード形式が依然としてサンプルデータに適合していることを確実にするために、サンプルデータに照らしてチェックしてもよい。

【００２９】

図４を参照すると、幾つかの実施態様では、システムは幾つかのサンプルレコードを含むサンプルデータを受け入れる。事前処理モジュール１０６は、データのレコード形式の識別を試行する。幾つかの実施態様では、既存の格納されたレコード形式に対する一致が見られない場合、データは、どのように符号化されるかを決定するために分析される。例えばデータは、ＡＳＣＩＩ又はＥＢＣＤＩＣ文字符号化に基づいて符号化するか、又は２進形式にしてもよい。幾つかの実施態様では、次にシステムは、システムが、データを解析することができる使用可能なパーザを有するか否かを決定する。システムは、サンプルデータに関するレコード形式を決定するために、サンプルデータを検査してもよい。例えば、テキストベースのサンプルデータは、区切りフィールド及びレコード、固定長フィールドを使用して形式化してもよく、拡張可能マークアップ言語（ＸＭＬ）などのタグ付きデータは標準汎用マークアップ言語（ＳＧＭＬ）である。データは、レコード形式の決定を支援するために、タグ又は区切り文字なしの２進形式としてもよい。２進データは、データベース、スプレッドシート、ワードプロセッシング文書、画像、又は他の２進データ

としてもよい。幾つかの実施態様では、２進データのデータタイプは、データ自体の検査に基づいて導出してもよい。他の実施態様では、２進データのデータタイプは、例えばファイル拡張子などのファイルの名前の一部に基づいて推測してもよい。システムは、サンプル形式の解析に基づいてフィールド及びレコードを決定してもよい。例えば、システムが区切りフィールド及びレコードを認識する場合、システムは区切り文字に基づいてデータをフィールド及びレコードに分離する。システムがタグ付きデータを認識する場合、システムはタグに基づいてファイルを解析する。

【００３０】

一例では、図４を参照すると、システムはサンプルデータファイル４０２を受信する。この例では、データはＡＳＣＩＩテキストを使用して符号化され、キャリッジリターンが異なるレコードを分離しているカンマ区切りフィールドを使用して構造化される。

【００３１】

プロセス矢印４０４によって表されるように、システムは、サンプルデータに関するレコード形式４０６を決定するためにサンプルデータの複数のレコードを分析する。この例でシステムは、文字列、文字列、ルックアップ値、電話番号、及び日付という５つのフィールドを識別する。整数、浮動小数点数、固定長テキストフィールド、及び固定長小数などの他のデータタイプも検出及び識別してもよい。幾つかの実施態様では、サンプルデータによって提供された値をプロファイリングすることによって、ルックアップフィールドに使用可能な値を識別してもよい。幾つかの実施態様では、サンプルデータのレコード形式が導出されると、システムはこのサンプルデータを解析して、それぞれのデータフィールドに対する値を決定してもよい。例えば、この情報を使用して、相対的に小さな数の有効値のみを含むフィールドを識別してもよい。幾つかの実施態様では、データのヒューリスティックスの分析に基づいて、サンプルデータのレコード形式を決定してもよい。例えば、固定長レコードのセットの長さはレコード数で均等に分けることも可能である。

【００３２】

幾つかの実施態様では、サンプルデータに関するレコード形式が決定されると、レコード形式はデータに関連付けられる。別の実施態様では、レコード形式をユーザに対して表示してもよく、ユーザはこのレコード形式を修正することができる。プロセス矢印４１４によって表されるように、修正されたレコード形式は、依然としてサンプルデータに適合していることを確認するために、サンプルデータに照らしてテストされる。レコード形式にサンプルデータの解析を実行不可にさせるデータタイプをユーザが入力した場合、システムはユーザにエラーを提示し、問題を訂正するレコード形式への変更を提案してもよい。この実施態様では、レコード形式は、レコード形式が最終的に確定されると、データに関連付けられる。

【００３３】

図５を参照すると、幾つかの実施態様では、システムは、ユーザによって提供可能であるか、又は本明細書で説明されるような検索技法を使用して識別してもよいサンプルデータ５０２と共に、可能なレコード形式５０４を受信する。可能なレコード形式がサンプルデータ５０２内のレコードの形式を正確に記述しているか否かについては、ある程度の不確定な要素がある。可能なレコード形式５０４は、ＸＭＬ文書タイプ定義、又はマスタからコピーすること、並びにＣＯＢＯＬコピーブック及びデータ操作言語（ＤＭＬ）レコード形式などの幾つかの異なるプログラムに挿入することができるプログラムデータの物理レイアウトを定義するコードのセクションとすることができる。

【００３４】

プロセス矢印５０８によって表されるように、システムは、処理中に発生するいずれかのエラーを示しながら、可能なレコード形式５０４を使用してサンプルデータの解析を試行する。この例では、第１のフィールドは可能なレコード形式では番号として定義されるが、サンプルデータ５０２では第１のフィールドは可変長文字フィールドである。システムがレコード形式を使用してデータの解析を試行した場合、エラーログ５０６が生成され、ユーザに提示される。ユーザには、不一致を解決するための提案が提供される。例えば

ユーザには、フィールド１のデータタイプを可変長文字フィールドに変更することの提案を提示してもよい。

【００３５】

図６を参照すると、幾つかの実施態様では、システムはサンプルデータ６０２を受信し、システムがデータを処理できるように、既存のレコード形式がデータ内のレコードの形式を正確に記述できるか否かを決定するように要求される。システムは、サンプルデータがレコード形式リポジトリ６０４内のいずれかの候補レコード形式６０６ａ～ｇと一致するか否かを決定するために、サンプルデータ内の複数のレコードを分析してもよい。幾つかの実施態様では、この分析は、レコード形式リポジトリ６０４内に格納された候補レコード形式６０６ａ～ｇのそれぞれを使用して、サンプルデータの解析を試行することを含んでもよい。幾つかの実施態様では、データの解析は、各レコード内の各フィールドのサンプル値を決定するために、候補レコード形式をサンプルデータに適用することを含む。サンプル値を候補レコード形式と比較して、サンプル値が候補レコード形式のそれらに適合するか否かを決定してもよい。幾つかの実施態様では、この分析は、候補レコード形式によってフィールドに対して確立された有効値又は有効値の範囲を定義する妥当性確認テストに照らして、サンプルデータ内の値の妥当性を確認することを含んでもよい。例えば、フィールドは、限定数の有効値（５０の州、２つの性別など）が可能である。

【００３６】

各レコード形式について、システムは妥当性確認テストと呼ばれる解析の成功の尺度を決定する。例えば１つの例示的妥当性確認テストでは、システムは、解析が成功しなかったレコード数のカウントを維持する。別の例示的妥当性確認テストでは、システムは、解析が成功しなかったフィールド数のカウント、並びにどのフィールドが処理できなかったかの指示を維持する。システムは、レコード形式を候補レコード形式セット、６０６ｅ、６０６ｆ、６０６ｇに縮小し、それらをユーザに提示する。幾つかの実施態様では、レコード形式は、サンプルデータに関連付けられたレコード形式と厳密に一致しない場合がある。例えば、候補レコード形式６０６ｅは最後が文字列フィールドであり、他の候補レコード形式は最後が日付フィールドであるが、文字列には日付値を入れることができるため、依然としてレコード形式はサンプルデータに適合している。他の解析の不整合も許容される場合がある。例えば１つのテストの場合、有効値の事前に定義された範囲外にある値が、依然として候補レコード形式を生成する場合があり、例えば、潜在的レコード形式６０６ｇは、有効値「Ｍ」及び「Ｓ」を伴う「配偶者の有無」フィールドを含む。サンプルデータセットは、「Ｍ」又は「Ｆ」のいずれかを含むフィールドを含む。システムは、解析エラーを示しながら、潜在的データレコード６０６ｇを含めることができる。幾つかのテストでは、解析エラーの数が所与のしきい値を下回る場合に、潜在的データレコードが含められる。他のテストでは、有効解析値が所与のしきい値を上回る場合に、潜在的データレコードが含められる。

【００３７】

幾つかの実施態様では、システムは候補レコード形式をユーザに提示し、ユーザがデータに適合するレコード形式を選択できるようにしてもよい。この例では、ユーザは、候補レコード形式６０６ｆを最良の適合として選択してもよい。幾つかの実施態様では、システムは適合可能なレコード形式を検査し、サンプルデータ及び候補レコード形式のプロファイルに基づいて、どのレコード形式が最良であるかを決定してもよい。幾つかの実施態様では、ユーザはレコード形式を修正してもよい。潜在的レコード形式のリストが単一のターゲットレコード形式に縮小されると、システムは、提供されたサンプルデータ６０２を解析することによって、選択されたターゲットレコード形式の妥当性を確認する。妥当性確認が完了した後、システムは、サンプルデータを選択されたターゲットレコード形式に関連付け、選択されたターゲットレコード形式を格納する、及び／又は選択されたターゲットレコード形式をユーザに提供する。幾つかの実施態様では、サンプルデータが、レコード形式内に提供されたデータタイプと合致しない場合、ユーザには、レコード形式をサンプルデータに合致させるようにレコード形式を修正するオプションが提示される。

【００３８】
幾つかの実施態様では、図７を参照すると、システムは、サンプルデータに適合するレコード形式リポジトリ６０４内の既存のレコード形式を識別することができない。このような状況下では、システムは、提供されたサンプルデータを既存のパーザが解析できるか否かを決定する。例えばサンプルデータセット７０２は、ＸＭＬ形式で示されている。この例では、レコード形式リポジトリ６０４は、サンプルデータに合致するいかなるレコード形式も含まない。プロセス矢印７０４によって示されるように、システムは、レコード形式がＸＭＬ形式のＡＳＣＩＩファイルであることを識別する。プロセス矢印７０８によって示されるように、システムは、既存のパーザ（例えばＸＭＬパーザ７１０）がデータの解釈を実行できるものと決定する。パーザ及びサンプルデータに基づいて、システムはサンプルデータ７１４のレコード形式を導出する。上述のように、システムは、パーザがサンプルデータを解釈できることを検証し、パーザによって生成された結果として生じるターゲットレコード形式をサンプルデータ７１４に関連付け、結果として生じるターゲットレコード形式をレコード形式リポジトリ内に格納する。幾つかの実施態様では、システムは、ターゲットレコード形式をレコード形式リポジトリ内に格納する前に、新しく作成されたターゲットレコード形式を承認のためにユーザに提示する。
【００３９】
図８は、ターゲットレコード形式を決定するために事前処理モジュール１０６が使用できる別の例示的プロセス８００に関するフローチャートを示す。事前処理モジュールの動作は、供給された入力データがサンプルデータを含むか否かの決定８０２を含む。
【００４０】
動作は、入力データがサンプルデータを含む場合に、サンプルデータをアップロード及び／又は位置付けすること８０４も含む。事前処理モジュールは、入力データによって定義された位置からサンプルにアクセスしてもよい。幾つかの実施態様では、事前処理モジュールは、別のサーバからアクセスポートを介してサンプルデータをアップロードするか又はこれにアクセスしてもよい。他の実施態様では、事前処理モジュールは、サンプルデータを含むファイル又は他のデータ記憶機構にアクセスしてもよい。
【００４１】
動作は、サンプルデータを分析すること８０６、及びオプションで分析結果を格納することも含む。サンプルデータは、文字セット、メタデータ、レコード形式タイプ、及び／又はレコード形式自体を決定するために分析してもよい。幾つかの実施態様では、システムは、レコード形式リポジトリ内に格納された１つ又は複数の既知のレコード形式の検索を実行するか否かを決定するために、サンプルデータを分析する。例えば事前処理モジュールは、サンプルデータが第１のタイプ（例えばカンマ区切りファイル）である場合は、潜在的レコード形式を決定するために検索を実行してもよいが、サンプルデータが第２のタイプ（例えばＸＭＬ）であることが決定された場合は、実行しない。他の実施態様では、レコード形式の作成及び妥当性確認を支援することができるメタデータを探索するために、サンプルデータが分析される。幾つかの実施態様では、事前処理モジュールは、フィールド区切り、エスケープ文字、及びフィールド名を含むヘッダを識別する。分析の結果は、後の意思決定プロセスで使用するために保持してもよい。
【００４２】
この実施態様では、動作は、サンプルデータを含む文書のタイプがＸＭＬであるか否かを決定すること８０８を含む。幾つかの実施態様では、この例におけるＸＭＬ形式などの１つ又は複数の所定の形式の文書は、他の形式の文書とは別に取り扱われる。この実施態様では、サンプルＸＭＬ文書はＸＭＬパーザによって処理される８２６。
【００４３】
動作は、サンプルデータが、レコード形式リポジトリ内に格納された１つ又は複数の既知のレコード形式と合致するか否かを決定すること８１０も含む。これは、上述のように、妥当性確認テストを使用して、サンプルデータ内の１つ又は複数のレコードの妥当性を各レコード形式に照らして確認すること、及び妥当性確認エラーの数を決定することによ

って達成できる。他の実施態様では、サンプルデータの分析８０６中に取得された情報を使用して、サンプルデータの妥当性が確認されたデータ形式の数を低減してもよい。

【００４４】

動作は、合致するレコード形式をユーザに示すこと８１２も含む。上述のように、事前処理モジュールは、潜在的に合致するレコード形式のリストをユーザに表示してもよい。

【００４５】

動作は、ユーザが、潜在的レコード形式のリストから合致するレコード形式を選択するか否かを決定すること８１４も含む。

【００４６】

動作は、格納されたレコード形式に対する合致が見つからない、及び／又はユーザによって選択されない場合、使用可能なパーザが存在する既知のネイティブ形式を有するファイルに含まれているなど、サンプルデータが既知のデータタイプを有するか否かを決定すること８１６も含む。ネイティブ形式とは、アプリケーション又はシステムによって使用される既知の外部形式である。

【００４７】

動作は、ネイティブ形式が既知の場合、適切な使用可能なパーザに対するデータの合致を決定することも含む。例えばサンプルデータは、既知のパーザによって処理することができるタグ付きレコードを含むことができる８２０。

【００４８】

動作は、タグ付きサンプルデータに関するパーザを識別すること８３０も含む。

【００４９】

この実施態様では、使用可能なパーザに対する合致を決定することは、サンプルデータがＣＯＢＯＬであるか否かを決定すること８２２を含む。幾つかの実施態様では、動作は、サンプルデータが、使用可能なパーザによって解析できる標準のデータレコード形式構造を使用する別のプログラミング言語であるか否かを決定することも含んでもよい。

【００５０】

動作は、サンプルデータがＣＯＢＯＬである場合、ＣＯＢＯＬコピーブックをアップロード及び解析すること８３２も含む。

【００５１】

既知のネイティブ形式と別の使用可能なパーザとを突き合せることは、サンプルデータがデータベース内に格納されていることを決定すること、及び事前処理モジュールがデータベースにアクセスしてもよいことの妥当性を確認すること８２４を含む。データベースへのアクセスは、事前処理モジュールが、例えばユーザ名及びパスワードなどの有効な信用証明にアクセスできるか否かを検証することを含んでもよい。データベースへのアクセスは、信用証明がサンプルデータへのアクセスを提供するか否かを決定することも含んでもよい。

【００５２】

動作は、データベース内に格納されたサンプルデータを分析すること、及びこの分析からレコード形式を決定すること（例えば、ＳＱＬエディタ内）８３４も含む。幾つかの実施態様では、事前処理モジュールは、レコード形式を導出するためにデータベースのテーブル構造を分析する。

【００５３】

既知のネイティブ形式と別の使用可能なパーザとを突き合せることは、サンプルデータがＸＭＬ形式であるか否か、及び文書タイプ定義又はＸＭＬスキーマ定義（ＸＳＤ）を含むか否かを決定すること８２６を含む。

【００５４】

動作は、ＸＭＬ文書の構造をレコード形式に変換すること（例えば、ＸＭＬ経路エディタ）８３６も含む。

【００５５】

既知のネイティブ形式と別の使用可能なパーザとを突き合せることは、データがＳＡＰ

形式であるか否かを決定すること８２８を含む。幾つかの実施態様では、他のエンタープライズソリューションソフトウェアパッケージが、例えば、Oracle Financialsからサンプルデータを検出してもよい。

【００５６】

動作は、エンタープライズソフトウェアパッケージに関するインポートモジュールを使用して、レコード形式を決定すること８３８も含む。

【００５７】

サンプルデータのデータタイプが知られていないか、又はデータタイプに使用可能なパーザがない場合、動作は、サンプルデータの特徴を決定すること、及びサンプルデータの特徴の分析から構築済みレコード形式を生成することを含む。例えばこの実施態様では、動作は、サンプルデータが大部分タグ付けされているか否かを決定すること８４０を含む。大部分タグ付けされたデータとは、例えば、主にタグ付きデータ構造を含むように見えるが、必ずしもタグ付き構造に適合しない幾つかのデータを含むデータである。

【００５８】

動作は、データが大部分タグ付けされたものと決定された場合、（例えばタグエディタを使用して）データをタグ付けデータとして処理するように試行すること８４２も含む。ＸＭＬに加えて、例えば国際銀行間金融通信協会（ＳＷＩＦＴ）形式などの他のタグ付き形式を取り扱ってもよい。

【００５９】

動作は、汎用タグ付きデータパーザ、又は既知のパーザが、サンプルデータを処理可能であるか否かを決定すること８４４も含む。

【００６０】

動作は、サンプルデータについてパーザビルダに照会すること８４８も含む。

【００６１】

動作は、サンプルデータが大部分テキストであるか否かを決定すること８５２も含む。大部分テキストのデータとは、例えば主に、例えばＡＳＣＩＩ又はＥＢＣＤＩＣなどの周知のテキスト形式を使用して符号化されたデータである。

【００６２】

動作は、データの構造を決定するよう試行すること８５４も含む。幾つかの実施態様では、データの構造は、レコード及びフィールド区切り文字を識別することによって決定してもよい。レコード区切り文字は、サンプルデータ内の最後の文字を検査することによって識別してもよい。区切り文字は、非印刷文字又は非英数文字に関するデータを検査することによっても識別してもよい。２つの非印刷可能文字又は非英数文字がサンプルデータ内に発生するケースでは、フィールド区切り文字が最も一般的であり、レコード区切り文字は一般的でない。区切り文字でない非印刷可能文字が存在することは、サンプルデータが区切られていないことを示してもよい。区切り文字を識別した後、事前処理モジュールはサンプルデータに区切り文字を適用し、不整合に関してチェックしてもよい。例えばシステムは、各レコードが同じ数のフィールドを含むか否かをチェックしてもよい。システムは、各レコード内の同じフィールドが、同様か又は適合するデータタイプを含むか否かをチェックしてもよい。幾つかの実施態様では、事前処理モジュールは、データの分析中８０６にデータに関して決定された情報に依拠する。

【００６３】

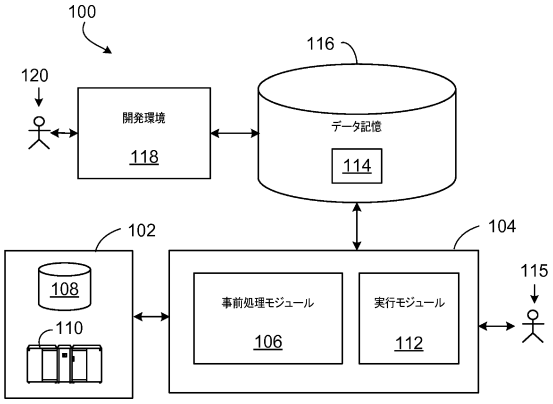
動作は、データが大部分２進であることを決定すること８５６も含む。２進データとは、例えば、ＡＳＣＩＩ及びＥＢＣＤＩＣなどの周知のテキスト形式を使用して符号化されていないデータである。

【００６４】

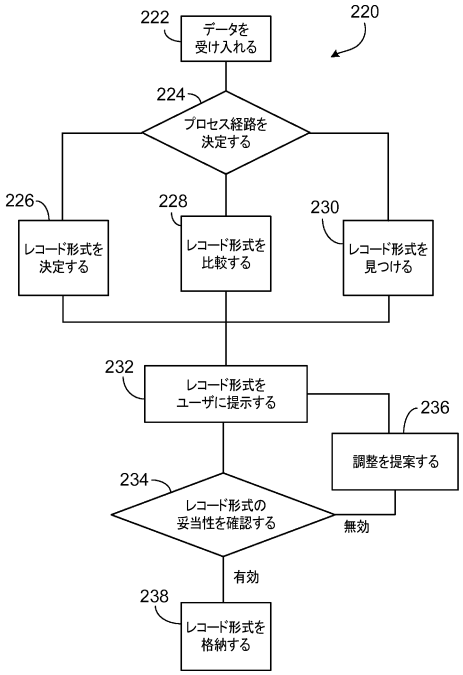
動作は、適切であれば（例えばデータが大部分２進であることの決定８５６に応答して）、フィールド名をサンプルデータに挿入すること８５８も含む。幾つかの実施態様では、ユーザは挿入することになるフィールド名を入力する（ペースト又はキー入力）ことができる。

【００６５】
動作は、結果を検証すること８５０も含む。レコード形式を検証することは、レコード形式を使用すること、及びサンプルデータの解析を試行することを含んでもよい。
【００６６】
動作は、ユーザがレコード形式を構築又は編集できるようにすること８４６も含む。幾つかの実施態様では、ユーザは、レコード形式を編集、及び／又はサンプルデータのタイプ、名前、及び構造を変更してもよい。
【００６７】
動作は、レコード形式リポジトリ内にレコード形式を格納すること８６０も含む。幾つかの実施態様では、事前処理モジュールはデータ形式をサンプルデータに関連付け、他の実施態様では、事前処理モジュールはデータ形式のコピーを作成し、そのコピーをデータに関連付ける。
【００６８】
上記レコード形式発見方法は、コンピュータ上で実行するためのソフトウェアを使用して実施することができる。例えばソフトウェアは、それぞれが少なくとも１つのプロセッサと、（揮発性及び不揮発性メモリ及び／又は記憶要素を含む）少なくとも１つのデータ記憶システムと、少なくとも１つの入力デバイス又はポートと、少なくとも１つの出力デバイス又はポートとを含む、（分散型、クライアント／サーバ、又はグリッドなどの様々なアーキテクチャであってよい）１つ又は複数のプログラム済み又はプログラマブルコンピュータシステム上で実行する、１つ又は複数のプログラム内に手順を形成する。ソフトウェアは、例えば計算グラフの設計及び構成に関する他のサービスを提供するより大きなプログラムの１つ又は複数のモジュールを形成してもよい。グラフのノード及び要素は、コンピュータ読み取り可能媒体に格納されたデータ構造、又はデータリポジトリに格納されたデータモデルに合致する他の編成データとして実施することができる。
【００６９】
ソフトウェアは、ＣＤ－ＲＯＭなどの記憶媒体上に提供できるか、汎用又は専用のプログラマブルコンピュータによって読み取り可能であるか、又は（伝搬される信号内で符号化されて）ネットワークの通信媒体を介して、実行されるコンピュータに送達されてもよい。すべての機能は、専用のコンピュータ上で、又はコプロセッサなどの専用のハードウェアを使用して実行されてもよい。ソフトウェアは、ソフトウェアによって指定された計算の異なる部分が異なるコンピュータによって実行される分散様式で実施されてもよい。このようなコンピュータプログラムのそれぞれは、本明細書で説明された手順を実行するために、記憶媒体又はデバイスがコンピュータシステムによって読み取られた場合に、コンピュータを構成及び動作させるために、好ましくは、汎用又は専用のプログラマブルコンピュータによって読み取り可能な記憶媒体又はデバイス（例えばソリッドステートメモリ又は媒体、或いは磁気又は光媒体）上に格納されるか又は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た、コンピュータ読み取り可能記憶媒体として実施されるものとみなすこともでき、そのように構成された記憶媒体は、本明細書で説明された機能を実行するために、特定及び事前に定義された様式でコンピュータシステムを動作させる。
【００７０】
本発明の幾つかの実施形態について説明してきた。それにもかかわらず、本発明の精神及び範囲を逸脱することなく、様々な修正を行うことができることを理解されよう。例えば、上記ステップのうちの幾つかは順序に依存していないため、記述された順序とは異なる順序で実行できる。
【００７１】
上記説明は、添付の特許請求の範囲によって定義された本発明の範囲を例示することを意図したものであり、これを限定するものでないことを理解されよう。例えば、上記幾つかの機能ステップは、処理全体に大幅に影響を与えることなく、異なる順序で実行することができる。他の実施形態は、以下の特許請求の範囲内にある。

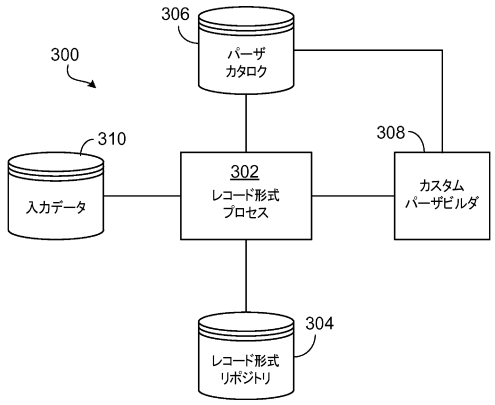
【図１】


100
116
120
開発環境
118
データ記憶
114
102
104
108
110
事前処理モジュール
106
実行モジュール
112
115

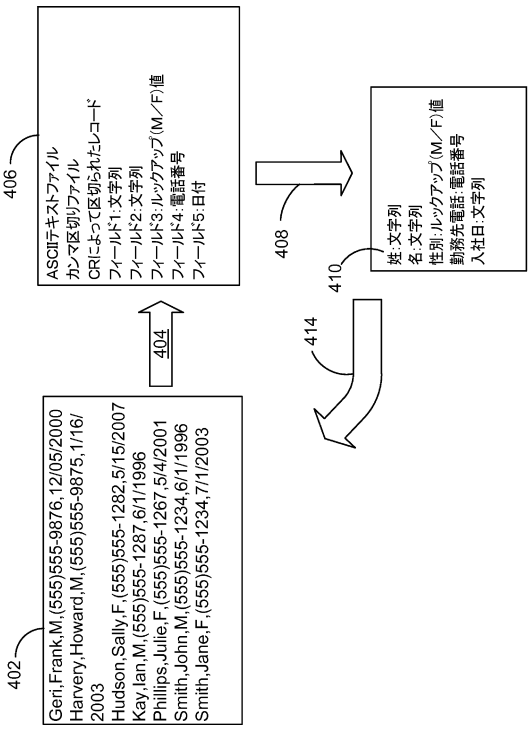
【図２】


222
データを
受け入れる
220
224
プロセス経路を
決定する
226
レコード形式を
決定する
228
レコード形式を
比較する
230
レコード形式を
見つける
232
レコード形式を
ユーザに提示する
236
調整を提案する
234
レコード形式の
妥当性を確認する
無効
有効
238
レコード形式を
格納する

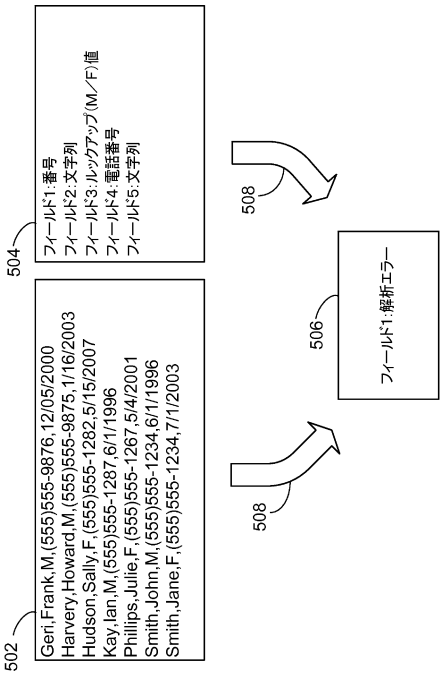
【図３】


306
パーザ
カタログ
300
310
308
302
レコード形式
プロセス
入力データ
カスタム
パーザビルダ
304
レコード形式
リポジトリ

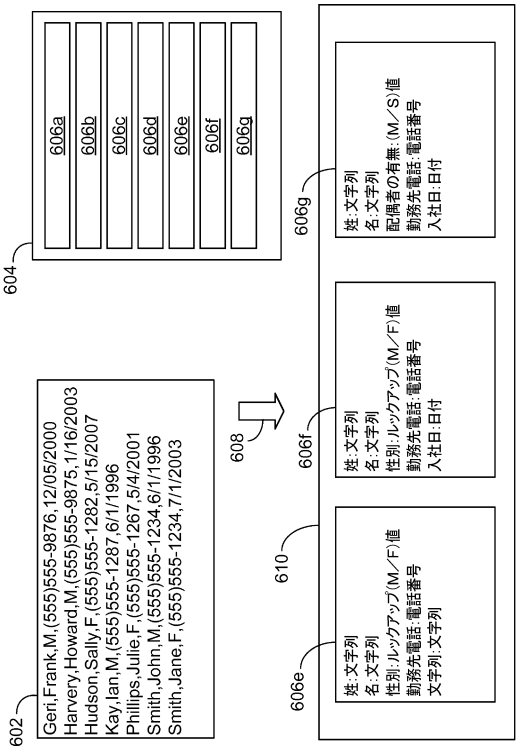
【図４】


406
ASCIIテキストファイル
カンマ区切りファイル
CRによって区切られたレコード
フィールド1:文字列
フィールド2:文字列
フィールド3:ルックアップ(M/F)値
フィールド4:電話番号
フィールド5:日付
408
410
姓:文字列
名:文字列
性別:ルックアップ(M/F)値
勤務先電話:電話番号
入社日:文字列
404
414
402
Geri,Frank,M,(555)555-9876,12/05/2000
Harvery,Howard,M,(555)555-9875,1/16/
2003
Hudson,Sally,F,(555)555-1282,5/15/2007
Kay,Ian,M,(555)555-1287,6/1/1996
Phillips,Julie,F,(555)555-1267,5/4/2001
Smith,John,M,(555)555-1234,6/1/1996
Smith,Jane,F,(555)555-1234,7/1/2003

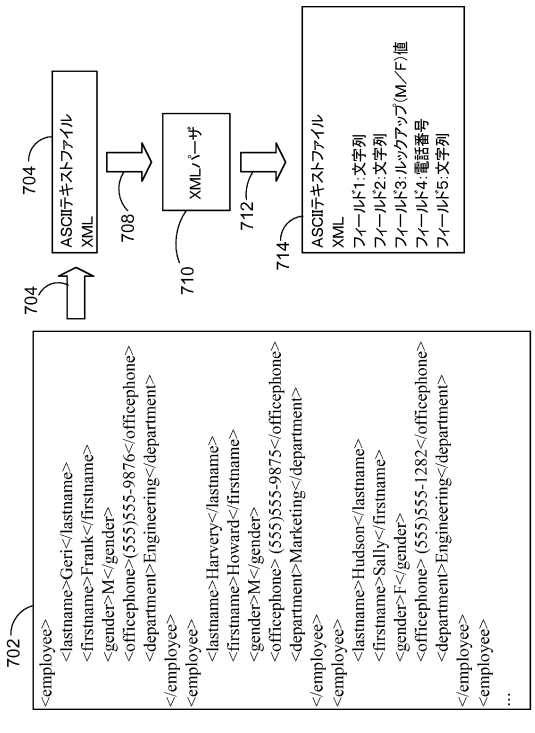
【図５】


502
Geri,Frank,M,(555)555-9876,12/05/2000
Harvery,Howard,M,(555)555-9875,1/16/2003
Hudson,Sally,F,(555)555-1282,5/15/2007
Kay,Ian,M,(555)555-1287,6/1/1996
Phillips,Julie,F,(555)555-1267,5/4/2001
Smith,John,M,(555)555-1234,6/1/1996
Smith,Jane,F,(555)555-1234,7/1/2003
504
フィールド1:番号
フィールド2:文字列
フィールド3:ルックアップ(M/F)値
フィールド4:電話番号
フィールド5:文字列
506
フィールド1:解析エラー
508
508

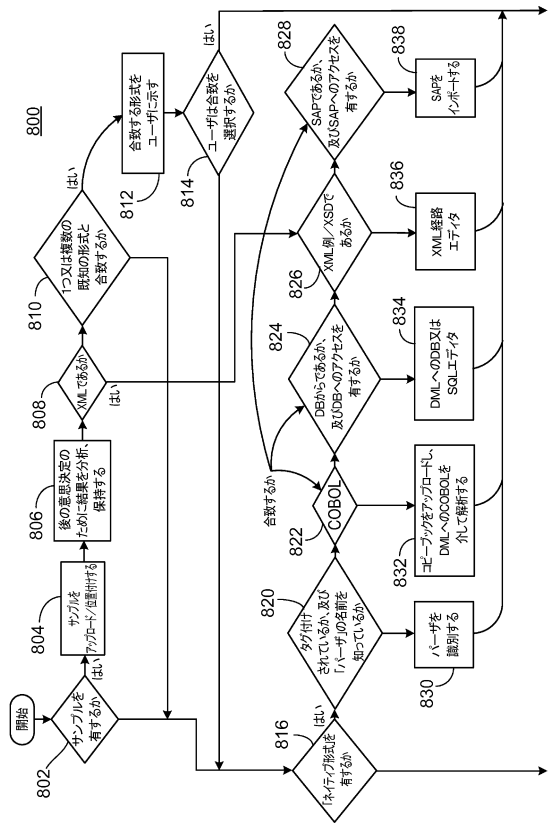
【図６】


602
Geri,Frank,M,(555)555-9876,12/05/2000
Harvery,Howard,M,(555)555-9875,1/16/2003
Hudson,Sally,F,(555)555-1282,5/15/2007
Kay,Ian,M,(555)555-1287,6/1/1996
Phillips,Julie,F,(555)555-1267,5/4/2001
Smith,John,M,(555)555-1234,6/1/1996
Smith,Jane,F,(555)555-1234,7/1/2003
604
606a
606b
606c
606d
606e
606f
606g
608
610
606e
姓:文字列
名:文字列
性別:ルックアップ(M/F)値
勤務先電話:電話番号
文字列:文字列
606f
姓:文字列
名:文字列
性別:ルックアップ(M/F)値
勤務先電話:電話番号
入社日:日付
606g
姓:文字列
名:文字列
配偶者の有無:(M/S)値
勤務先電話:電話番号
入社日:日付

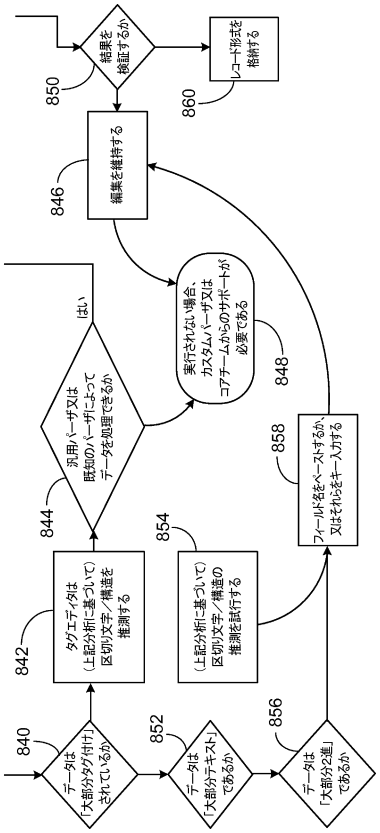
【図７】


702
<employee>
<lastname>Geri</lastname>
<firstname>Frank</firstname>
<gender>M</gender>
<officephone>(555)555-9876</officephone>
<department>Engineering</department>
</employee>
<employee>
<lastname>Harvery</lastname>
<firstname>Howard</firstname>
<gender>M</gender>
<officephone> (555)555-9875</officephone>
<department>Marketing</department>
</employee>
<employee>
<lastname>Hudson</lastname>
<firstname>Sally</firstname>
<gender>F</gender>
<officephone> (555)555-1282</officephone>
<department>Engineering</department>
</employee>
<employee>
...
704
704
ASCIIテキストファイル
XML
708
710
XMLパーザ
712
714
ASCIIテキストファイル
XML
フィールド1:文字列
フィールド2:文字列
フィールド3:ルックアップ(M/F)値
フィールド4:電話番号
フィールド5:文字列

【図８Ａ】


800
開始
802
サンプルを有するか
はい
804
サンプルをアップロード/値割付する
806
後の意思決定のために結果を分析、保持する
808
XMLであるか
はい
810
1つ又は複数の既知の形式と合致するか
はい
812
合致する形式をユーザに示す
814
ユーザは合致を選択するか
はい
816
ネイティブ形式を有するか
はい
820
タグ付けされているか、及び「パーザの名前を知っているか
822
COBOL
合致するか
824
DBからであるか、及びDB へのアクセスを有するか
826
XML例/XSDであるか
828
SAPであるか、及びSAPへのアクセスを有するか
830
パーザを識別する
832
コピーブックをアップロードし、DML へのCOBOLを介して解析する
834
DMLへのDB又はSQLエディタ
836
XML経路エディタ
838
SAPをインポートする

【図８Ｂ】


840
データは「大部分がタグ付け」されているか
842
タグエディタは（上記分析に基づいて）区切り文字／構造を推測する
844
汎用パーザ又は既知のパーザによってデータを処理できるか
はい
846
編集を維持する
848
実行されない場合カスタムパーザ又はコアチームからのサポートが必要である
850
結果を検証するか
860
レコード形式を格納する
852
データは「大部分テキスト」であるか
854
（上記分析に基づいて）区切り文字／構造の推測を試行する
856
データは「大部分2進」であるか
858
フィールド名をペーストするか、又はそれらをキー入力する

フロントページの続き

(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
(72)発明者　ファーバー，ジェニファー，エム．
アメリカ合衆国，イリノイ州　６０６１５，シカゴ，エス．ケンウッド　アベニュー　４８１５
(72)発明者　フロイントリッヒ，ロバート
アメリカ合衆国，マサチューセッツ州　０１７７６，サッドバリー，メープル　アベニュー　５５
(72)発明者　ヴィグノー，ジョイス，エル．
アメリカ合衆国，マサチューセッツ州　０２１８０，ストーンハム，フォレスト　ストリート　４５

審査官　池田　聡史

(56)参考文献　特開２００１－１０１０４９（ＪＰ，Ａ）
国際公開第２００６／０４６６６５（WO，Ａ１）
特開２００８－３０５３４８（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　１２／００