(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第5525541号
(P5525541)
(45)発行日　平成26年6月18日(2014.6.18)　(24)登録日　平成26年4月18日(2014.4.18)

(51)Int.Cl.　　F I
G06T 1/00 (2006.01)　　G06T 1/00　B
G06F 12/00 (2006.01)　　G06F 12/00　517

請求項の数 98　(全 25 頁)

(21)出願番号　特願2011-539631 (P2011-539631)
(86)(22)出願日　平成21年12月1日 (2009.12.1)
(65)公表番号　特表2012-510687 (P2012-510687A)
(43)公表日　平成24年5月10日 (2012.5.10)
(86)国際出願番号　PCT/US2009/066210
(87)国際公開番号　WO2010/065511
(87)国際公開日　平成22年6月10日 (2010.6.10)
審査請求日　平成24年11月30日 (2012.11.30)
(31)優先権主張番号　61/119,164
(32)優先日　平成20年12月2日 (2008.12.2)
(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　ウェイクリング　ティム
アメリカ合衆国　０１８１０　マサチューセッツ州　アンドーヴァー　アボット　ストリート　１１

最終頁に続く

(54)【発明の名称】データ管理システム内のデータセットのインスタンスのマッピング

(57)【特許請求の範囲】
【請求項１】
少なくとも一つの入力データセットからデータを受取り、且つ、少なくとも一つの出力データセットにデータを出力するデータ処理プログラムの仕様を処理するステップと、
第１のデータセット候補を特定するステップであって、前記第１のデータセット候補は前記データ処理プログラムのうちの第１のデータ処理プログラムに表現され、前記第１のデータセット候補は第１のデータセットの異なるバージョンを特定するための一つ以上の基準に一致する、ステップと、
第２のデータセット候補を特定するステップであって、前記第２のデータセット候補は前記データ処理プログラムのうちの第２のデータ処理プログラムに表現され、前記第２のデータセット候補は前記第１のデータセットの異なるバージョンを特定するための一つ以上の基準に一致する、ステップと、
前記第１のデータセット候補と前記第２のデータセット候補との間のマッピングを受け取るユーザインターフェースを提供するステップであって、前記マッピングは、前記第１のデータセット候補と前記第２のデータセット候補の双方が同じデータセットを表現することを示す、ステップと、
前記マッピングの前記第１のデータセット候補もしくは前記第２のデータセット候補に対してデータを提供する、または、前記マッピングの前記第１のデータセット候補もしくは前記第２のデータセット候補からデータを受け取るデータ処理プログラムに関連して前記ユーザインターフェースを通して受け取られた前記マッピングを格納するステップと、

を含む方法。
【請求項２】
前記セットを前記ユーザインターフェースを通して提示することを含む請求項１に記載の方法。
【請求項３】
前記一つ以上の基準への一致の数量化に従って順位付けられた、可能性のあるマッピングのリストを前記ユーザインターフェースを通して提示することを含む請求項１に記載の方法。
【請求項４】
前記可能性のあるマッピングのリストは、前記リスト内にてより高位に順序付けられた所与のデータセットのバージョンである可能性がより高い候補を含む請求項３に記載の方法。
【請求項５】
前記基準の一つが前記第１のデータセット候補及び前記第２のデータセット候補を特定するマッパに組込まれている請求項３に記載の方法。
【請求項６】
前記基準の一つは前記ユーザインターフェースから受け取られる請求項３に記載の方法。
【請求項７】
前記可能性のあるマッピングの少なくとも一つは前記データセット候補の一つを表現するデータフローグラフの構成要素を示し、且つ、前記可能性のあるマッピングの少なくとも一つは前記データセット候補の一つを表現しないデータフローグラフの構成要素を示す、請求項３に記載の方法。
【請求項８】
複数の構成要素を含むデータフローグラフのサブグラフは前記データセット候補の一つを表現する請求項７に記載の方法。
【請求項９】
前記サブグラフはデータ構成要素を含む請求項８に記載の方法。
【請求項１０】
前記サブグラフは実行可能な構成要素を含む請求項８に記載の方法。
【請求項１１】
前記第１のデータセット候補及び前記第２のデータセット候補を特定することは、前記第１のデータセット候補が前記第２のデータセット候補と共通の一つ以上の特徴を有するか否かを判別するためのヒューリスティックスを使用することを含む請求項１に記載の方法。
【請求項１２】
前記特徴は前記データセット候補の一つの表現におけるバイト及びレコードの量を含む請求項１１に記載の方法。
【請求項１３】
前記特徴は前記データセット候補の一つの表現の名称を含む請求項１１に記載の方法。
【請求項１４】
前記特徴は前記データセット候補の一つの表現の生成日を含む請求項１１に記載の方法。
【請求項１５】
前記特徴は前記データセット候補の一つの表現のデータフォーマットを含む請求項１１に記載の方法。
【請求項１６】
前記マッピングの前記第１のデータセット候補及び前記第２のデータセット候補の少なくとも一つはデータ管理システムに知られているデータセットのグループに属する請求項１に記載の方法。

【請求項１７】
さらに、前記第１のデータセット候補及び前記第２のデータセット候補の間のフォーマットマッピングを提供することを含む請求項１に記載の方法。
【請求項１８】
前記マッピングは、前記第１のデータセット候補及び前記第２のデータセット候補を追跡する前記データ管理システムにおけるレコードを指し示す識別子を含む請求項１に記載の方法。
【請求項１９】
さらに、前記第１のデータセット候補及び前記第２のデータセット候補の一方又は双方の変化に基づいて前記マッピングを更新することを含む請求項１に記載の方法。
【請求項２０】
データストレージシステム内に格納されたデータをマッピングするシステムであって、
命令を格納するメモリと、
前記格納された命令を実行するプロセッサと、
少なくとも一つの入力データセットからデータを受取り、且つ、少なくとも一つの出力データセットにデータを出力するデータ処理プログラムの仕様を格納するデータストレージシステムと、
第１のデータセット候補を特定し、前記第１のデータセット候補は前記データ処理プログラムのうちの第１のデータ処理プログラムに表現され、前記第１のデータセット候補は第１のデータセットの異なるバージョンを特定するための一つ以上の基準に一致する、マッパであって、
第２のデータセット候補を特定し、前記第２のデータセット候補は前記データ処理プログラムのうちの第２のデータ処理プログラムに表現され、前記第２のデータセット候補は前記第１のデータセットの異なるバージョンを特定するための一つ以上の基準に一致する、マッパと、
前記第１のデータセット候補と前記第２のデータセット候補との間のマッピングを受け取るユーザインターフェースであって、前記マッピングは、前記第１のデータセット候補と前記第２のデータセット候補の双方が同じデータセットを表現することを示し、前記ユーザインターフェースは、前記マッピングの前記第１のデータセット候補もしくは前記第２のデータセット候補に対してデータを提供する、または、前記マッピングの前記第１のデータセット候補もしくは前記第２のデータセット候補からデータを受け取るデータ処理プログラムに関連して前記データストレージシステム内の前記マッピングを格納する、ユーザインターフェースと
を含むシステム。
【請求項２１】
前記ユーザインターフェースが前記セットを提示する請求項２０に記載のシステム。
【請求項２２】
前記ユーザインターフェースが、前記一つ以上の基準への一致の数量化に従って順位付けられた、可能性のあるマッピングのリストを提示する請求項２０に記載のシステム。
【請求項２３】
前記可能性のあるマッピングのリストは、前記リスト内にてより高位に順序付けられた所与のデータセットのバージョンである可能性がより高い候補を含む請求項２２に記載のシステム。
【請求項２４】
前記基準の一つが前記マッパに組み込まれている請求項２２に記載のシステム。
【請求項２５】
前記基準の一つは前記ユーザインターフェースによって受け取られる請求項２２に記載のシステム。
【請求項２６】
前記可能性のあるマッピングの少なくとも一つは前記データセット候補の一つを表現す

るデータフローグラフの構成要素を示し、且つ、前記可能性のあるマッピングの少なくとも一つは前記データセット候補の一つを表現しないデータフローグラフの構成要素を示す、請求項２２に記載のシステム。

【請求項２７】

複数の構成要素を含むデータフローグラフのサブグラフは前記データセット候補の一つを表現する請求項２６に記載のシステム。

【請求項２８】

前記サブグラフはデータ構成要素を含む請求項２７に記載のシステム。

【請求項２９】

前記サブグラフは実行可能な構成要素を含む請求項２７に記載のシステム。

【請求項３０】

前記マッパは、前記第１のデータセット候補が前記第２のデータセット候補と共通の一つ以上の特徴を有するか否かを判別するためのヒューリスティックスを使用する請求項２０に記載のシステム。

【請求項３１】

前記特徴は前記データセット候補の一つの表現におけるバイト及びレコードの量を含む請求項３０に記載のシステム。

【請求項３２】

前記特徴は前記データセット候補の一つの表現の名称を含む請求項３０に記載のシステム。

【請求項３３】

前記特徴は前記データセット候補の一つの表現の生成日を含む請求項３０に記載のシステム。

【請求項３４】

前記特徴は前記データセット候補の一つの表現のデータフォーマットを含む請求項３０に記載のシステム。

【請求項３５】

前記マッピングの前記第１のデータセット候補及び前記第２のデータセット候補の少なくとも一つはデータ管理システムに知られているデータセットのグループに属する請求項２０に記載のシステム。

【請求項３６】

前記マッパは、前記第１のデータセット候補と前記第２のデータセット候補との間のフォーマットマッピングを発生する請求項２０に記載のシステム。

【請求項３７】

前記マッピングは、前記前記第１のデータセット候補及び前記第２のデータセット候補を追跡する前記データ管理システムにおけるレコードを指し示す識別子を含む請求項２０に記載のシステム。

【請求項３８】

前記マッパは前記第１のデータセット候補及び前記第２のデータセット候補の一方又は双方の変化に基づいて前記マッピングを更新する請求項２０に記載のシステム。

【請求項３９】

データストレージシステムに格納されたデータをマッピングするシステムであって、

少なくとも一つの入力データセットからデータを受取り、且つ、少なくとも一つの出力データセットにデータを出力するデータ処理プログラムの仕様を格納するデータストレージシステムと、

第１のデータセット候補を特定し、前記第１のデータセット候補は前記データ処理プログラムのうちの第１のデータ処理プログラムに表現され、前記第１のデータセット候補は第１のデータセットの異なるバージョンを特定するための一つ以上の基準に一致する、手段であって、

第２のデータセット候補を特定し、前記第２のデータセット候補は前記データ処理プロ

グラムのうちの第２のデータ処理プログラムに表現され、前記第２のデータセット候補は前記第１のデータセットの異なるバージョンを特定するための一つ以上の基準に一致する、手段と、

前記第１のデータセット候補と前記第２のデータセット候補との間のマッピングを受け取るユーザインターフェースを提供する手段であって、前記マッピングは、前記第１のデータセット候補と前記第２のデータセット候補の双方が同じデータセットを表現することを示す、手段と、

前記マッピングの前記第１のデータセット候補もしくは前記第２のデータセット候補に対してデータを提供する、または前記マッピングの前記第１のデータセット候補もしくは前記第２のデータセット候補からデータを受け取るデータ処理プログラムに関連して前記ユーザインターフェースを通して受け取られた前記マッピングを格納する手段と、

を含むシステム。

【請求項４０】

データストレージシステムに格納されたデータをマッピング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ストレージデバイスであって、前記コンピュータプログラムは、コンピュータに、

少なくとも一つの入力データセットからデータを受取り、且つ、少なくとも一つの出力データセットにデータを出力するデータ処理プログラムの仕様を処理させる命令と、

第１のデータセット候補を特定させる命令であって、前記第１のデータセット候補は前記データ処理プログラムのうちの第１のデータ処理プログラムに表現され、前記第１のデータセット候補は第１のデータセットの異なるバージョンを特定するための一つ以上の基準に一致する、命令と、

第２のデータセット候補を特定させる命令であって、前記第２のデータセット候補は前記データ処理プログラムのうちの第２のデータ処理プログラムに表現され、前記第２のデータセット候補は前記第１のデータセットの異なるバージョンを特定するための一つ以上の基準に一致する、命令と、

前記第１のデータセット候補と前記第２のデータセット候補との間のマッピングを受け取るユーザインターフェースを提供させる命令であって、前記マッピングは、前記第１のデータセット候補と前記第２のデータセット候補の双方が同じデータセットを表現することを示す、命令と、

前記マッピングの前記第１のデータセット候補もしくは前記第２のデータセット候補に対してデータを提供する、または前記マッピングの前記第１のデータセット候補もしくは前記第２のデータセット候補からデータを受け取るデータ処理プログラムに関連して前記ユーザインターフェースを通して受け取られた前記マッピングを格納させる命令と

を含む、コンピュータ読み取り可能ストレージデバイス。

【請求項４１】

前記第１のデータセットの各バージョンは、前記データストレージシステムに関連した異なる場所に格納される請求項１に記載の方法。

【請求項４２】

前記第１のデータセットの各バージョンは、異なるデータストレージフォーマットを用いて解釈される請求項１に記載の方法。

【請求項４３】

前記第１のデータセットの各バージョンは、前記データ処理プログラムの実行間で変化するパラメータを用いてアクセスされる請求項１に記載の方法。

【請求項４４】

前記第１のデータセットの各バージョンは、前記データストレージシステムに関連した異なる場所に格納される請求項２０に記載のシステム。

【請求項４５】

前記第１のデータセットの各バージョンは、異なるデータストレージフォーマットを用いて解釈される請求項２０に記載のシステム。

【請求項４６】
前記第１のデータセットの各バージョンは、前記データフローグラフの実行間で変化するパラメータを用いてアクセスされる請求項２０に記載のシステム。
【請求項４７】
前記ユーザインターフェースが前記セットを提示する請求項３９に記載のシステム。
【請求項４８】
前記ユーザインターフェースが、前記一つ以上の基準への一致の数量化に従って順位付けられた、可能性のあるマッピングのリストを提示する請求項３９に記載のシステム。
【請求項４９】
前記可能性のあるマッピングのリストは、前記リスト内にてより高位に順序付けられた所与のデータセットのバージョンである可能性がより高い候補を含む請求項４８に記載のシステム。
【請求項５０】
前記第１のデータセット候補及び前記第２のデータセット候補を特定する前記手段は、前記基準の一つを含む請求項４８に記載のシステム。
【請求項５１】
前記基準の一つは、前記ユーザインターフェースによって受信される請求項４８に記載のシステム。
【請求項５２】
前記可能性のあるマッピングの少なくとも一つは前記データセット候補の１つを表現するデータフローグラフの構成要素を示し、且つ、前記可能性のあるマッピングの少なくとも一つは前記データセット候補の１つを表現しないデータフローグラフの構成要素を示す、請求項４８に記載のシステム。
【請求項５３】
複数の構成要素を含むデータフローグラフのサブグラフは前記データセット候補の１つを表現する請求項５２に記載のシステム。
【請求項５４】
前記サブグラフはデータ構成要素を含む請求項５３に記載のシステム。
【請求項５５】
前記サブグラフは実行可能な構成要素を含む請求項５３に記載のシステム。
【請求項５６】
前記第１のデータセット候補及び前記第２のデータセット候補を特定する前記手段は、前記第１のデータセット候補が前記第２のデータセット候補と共通の一つ以上の特徴を有するか否かを判別するためのヒューリスティックスを使用する請求項３９に記載のシステム。
【請求項５７】
前記特徴は前記データセット候補の一つの表現におけるバイト及びレコードの量を含む請求項５６に記載のシステム。
【請求項５８】
前記特徴は前記データセット候補の一つの表現の名称を含む請求項５６に記載のシステム。
【請求項５９】
前記特徴は前記データセット候補の一つの表現の生成日を含む請求項５６に記載のシステム。
【請求項６０】
前記特徴は前記データセット候補の一つの表現のデータフォーマットを含む請求項５６に記載のシステム。
【請求項６１】
前記マッピングの前記第１のデータセット候補及び前記第２のデータセット候補の少なくとも一つはデータ管理システムに知られているデータセットのグループに属する請求項

３９に記載のシステム。
【請求項６２】
前記第１のデータセット候補及び前記第２のデータセット候補を特定する前記手段は、前記第１のデータセット候補と前記第２のデータセット候補との間のフォーマットマッピングを発生する請求項３９に記載のシステム。
【請求項６３】
前記第１のデータセットの各バージョンは、前記データストレージシステムに関連した異なる場所に格納される請求項３９に記載のシステム。
【請求項６４】
前記第１のデータセットの各バージョンは、異なるデータストレージフォーマットを用いて解釈される請求項３９に記載のシステム。
【請求項６５】
前記第１のデータセットの各バージョンは、前記データフローグラフの実行間で変化するパラメータを用いてアクセスされる請求項３９に記載のシステム。
【請求項６６】
前記コンピュータプログラムは、コンピュータに、前記第１のデータセット候補及び前記第２のデータセット候補を前記ユーザインターフェースを通して提示させる命令をさらに含む、請求項４０に記載のコンピュータ読み取り可能ストレージデバイス。
【請求項６７】
前記コンピュータプログラムは、コンピュータに、前記一つ以上の基準への一致の数量化に従って順位付けられた、可能性のあるマッピングのリストを前記ユーザインターフェースを通して提示させる命令をさらに含む、請求項４０に記載のコンピュータ読み取り可能ストレージデバイス。
【請求項６８】
前記可能性のあるマッピングのリストは、前記リスト内にてより高位に順序付けられた所与のデータセットのバージョンである可能性がより高い候補を含む請求項６７に記載のコンピュータ読み取り可能ストレージデバイス。
【請求項６９】
前記基準の一つが前記第１のデータセット候補及び前記第２のデータセット候補を特定するマッパに組込まれている請求項６７に記載のコンピュータ読み取り可能ストレージデバイス。
【請求項７０】
前記基準の一つは前記ユーザインターフェースから受け取られる請求項６７に記載のコンピュータ読み取り可能ストレージデバイス。
【請求項７１】
前記可能性のあるマッピングの少なくとも一つは前記データセット候補の一つを表現するデータフローグラフの構成要素を示し、且つ、前記可能性のあるマッピングの少なくとも一つは前記データセット候補の一つを表現しないデータフローグラフの構成要素を示す、請求項６７に記載のコンピュータ読み取り可能ストレージデバイス。
【請求項７２】
複数の構成要素を含むデータフローグラフのサブグラフは前記データセット候補の一つを表現する請求項７１に記載のコンピュータ読み取り可能ストレージデバイス。
【請求項７３】
前記サブグラフはデータ構成要素を含む請求項７２に記載のコンピュータ読み取り可能ストレージデバイス。
【請求項７４】
前記サブグラフは実行可能な構成要素を含む請求項７２に記載のコンピュータ読み取り可能ストレージデバイス。
【請求項７５】
前記第１のデータセット候補及び前記第２のデータセット候補を特定することは、前記

第１のデータセット候補が前記第２のデータセット候補と共通の一つ以上の特徴を有するか否かを判別するためのヒューリスティックスを使用することを含む請求項４０に記載のコンピュータ読み取り可能ストレージデバイス。

【請求項７６】

前記特徴は前記データセット候補の一つの表現におけるバイト及びレコードの量を含む請求項７５に記載のコンピュータ読み取り可能ストレージデバイス。

【請求項７７】

前記特徴は前記データセット候補の一つの表現の名称を含む請求項７５に記載のコンピュータ読み取り可能ストレージデバイス。

【請求項７８】

前記特徴は前記データセット候補の一つの表現の生成日を含む請求項７５に記載のコンピュータ読み取り可能ストレージデバイス。

【請求項７９】

前記特徴は前記データセット候補の一つの表現のデータフォーマットを含む請求項７５に記載のコンピュータ読み取り可能ストレージデバイス。

【請求項８０】

前記マッピングの前記第１のデータセット候補及び前記第２のデータセット候補の少なくとも一つはデータ管理システムに知られているデータセットのグループに属する請求項４０に記載のコンピュータ読み取り可能ストレージデバイス。

【請求項８１】

前記コンピュータプログラムは、コンピュータに、前記第１のデータセット候補と前記第２のデータセット候補との間のフォーマットマッピングを提供させる命令をさらに含む、請求項４０に記載のコンピュータ読み取り可能ストレージデバイス。

【請求項８２】

前記マッピングは、前記第１のデータセット候補及び前記第２のデータセット候補を追跡するデータ管理システムにおけるレコードを指し示す識別子を含む請求項４０に記載のコンピュータ読み取り可能ストレージデバイス。

【請求項８３】

前記コンピュータプログラムは、コンピュータに、前記第１のデータセット候補及び前記第２のデータセット候補の一方又は双方の変化に基づいて前記マッピングを更新させる命令をさらに含む、請求項４０に記載のコンピュータ読み取り可能ストレージデバイス。

【請求項８４】

前記第１のデータセットの各バージョンは、前記データストレージシステムに関連した異なる場所に格納される請求項４０に記載のコンピュータ読み取り可能ストレージデバイス。

【請求項８５】

前記第１のデータセットの各バージョンは、異なるデータストレージフォーマットを用いて解釈される請求項４０に記載のコンピュータ読み取り可能ストレージデバイス。

【請求項８６】

前記第１のデータセットの各バージョンは、前記データフローグラフの実行間で変化するパラメータを用いてアクセスされる請求項４０に記載のコンピュータ読み取り可能ストレージデバイス。

【請求項８７】

前記データ処理プログラムの少なくとも一つは、データのフローを表現するリンクによって相互接続された計算を表現するノードを含むデータフローグラフを含み、前記データフローグラフは、前記少なくとも一つの入力データセットからデータのフローを受け取り、且つ、前記少なくとも一つの出力データセットにデータのフローを提供する請求項１に記載の方法。

【請求項８８】

前記第１のデータ処理プログラムはデータフローグラフのサブグラフであり、前記第２

のデータ処理プログラムは前記データフローグラフのサブグラフである、請求項１に記載の方法。

【請求項８９】

前記データ処理プログラムの少なくとも一つは、データのフローを表現するリンクによって相互接続された計算を表現するノードを含むデータフローグラフを含み、前記データフローグラフは、前記少なくとも一つの入力データセットからデータのフローを受け取り、且つ、前記少なくとも一つの出力データセットにデータのフローを提供する請求項２０に記載のシステム。

【請求項９０】

前記第１のデータ処理プログラムはデータフローグラフのサブグラフであり、前記第２のデータ処理プログラムは前記データフローグラフのサブグラフである、請求項２０に記載のシステム。

【請求項９１】

前記データ処理プログラムの少なくとも一つは、データのフローを表現するリンクによって相互接続された計算を表現するノードを含むデータフローグラフを含み、前記データフローグラフは、前記少なくとも一つの入力データセットからデータのフローを受け取り、且つ、前記少なくとも一つの出力データセットにデータのフローを提供する請求項３９に記載のシステム。

【請求項９２】

前記第１のデータ処理プログラムはデータフローグラフのサブグラフであり、前記第２のデータ処理プログラムは前記データフローグラフのサブグラフである、請求項３９に記載のシステム。

【請求項９３】

前記データ処理プログラムの少なくとも一つは、データのフローを表現するリンクによって相互接続された計算を表現するノードを含むデータフローグラフを含み、前記データフローグラフは、前記少なくとも一つの入力データセットからデータのフローを受け取り、且つ、前記少なくとも一つの出力データセットにデータのフローを提供する請求項４０に記載のコンピュータ読み取り可能ストレージデバイス。

【請求項９４】

前記第１のデータ処理プログラムはデータフローグラフのサブグラフであり、前記第２のデータ処理プログラムは前記データフローグラフのサブグラフである、請求項４０に記載のコンピュータ読み取り可能ストレージデバイス。

【請求項９５】

前記第１のデータセットの各バージョンは、異なるデータフローグラフ、データフローグラフサブセットまたは実行可能な構成要素に関連する請求項８７に記載の方法。

【請求項９６】

前記第１のデータセットの各バージョンは、異なるデータフローグラフ、データフローグラフサブセットまたは実行可能な構成要素に関連する請求項８９に記載のシステム。

【請求項９７】

前記第１のデータセットの各バージョンは、異なるデータフローグラフ、データフローグラフサブセットまたは実行可能な構成要素に関連する請求項９１に記載のシステム。

【請求項９８】

前記第１のデータセットの各バージョンは、異なるデータフローグラフ、データフローグラフサブセットまたは実行可能な構成要素に関連する請求項９３に記載のコンピュータ読み取り可能ストレージデバイス。

【発明の詳細な説明】

【技術分野】

【０００１】

本発明はデータ管理システム内のデータセット（データセット）のインスタンス（instance）のマッピングに関する。

【背景技術】
【０００２】
近年のデータ管理システムは、そのシステムの異なる特徴を表す多数の要素を含む。より複雑でないシステムは、しばしば、正確な視覚化の目的のために追加の処理なしでデータを直接見られることを可能にする。より複雑なシステムは、意味あるようにデータを見ることができるようにするために追加のメカニズムを必要とするであろう。多くの要素よりなる複雑なデータ管理システムは、データを多く異なる形式で格納し、且つ、データを多くの異なる方法で処理することができる。これらの格納形式及び処理形式の多くは、相互関係を解析する方法なしでは明白とならない態様において相互に関係する。
【発明の概要】
【０００３】
＜関連出願＞
本願は、２００８年１２月２日出願の米国特許出願第６１／１１９，１６４号の優先権を主張するものであり、ここに引用して組み込まれる。
概して、コンピュータシステムによって使用される、データストレージシステム（データ格納システム）内に格納されたデータをマッピングする方法は、
ノードを含むデータフローグラフの仕様を処理するステップであって、前記ノードがデータのフローを表現するリンクによって相互接続された計算を表現し、前記データフローグラフの少なくとも一つが少なくとも一つの入力データセットからデータのフローを受取り、且つ、前記データフローグラフの少なくとも一つが少なくとも一つの出力データセットにデータのフローを提供する、ステップ；
データセットの一つ以上のセットを特定するステップであって、所与のセット内の各データセットが単一のデータセットの異なるバージョンを特定するための一つ以上の基準に一致する、ステップ；
所与のセットにおける少なくとも二つのデータセットの間のマッピングを受け取るユーザインターフェースを提供するステップ；及び、
データを前記マッピングのデータセットに提供するかまたはデータを前記マッピングのデータセットから受け取る一つのデータフローグラフに関連して前記ユーザインターフェースを通して受け取られた前記マッピングを格納するステップ；
を含む。
【０００４】
他の概要において、データストレージシステム内に格納されたデータをマッピングするシステムは、
ノードを含むデータフローグラフの仕様を格納するデータストレージシステム（データ格納システム）であって、前記ノードがデータのフローを表現するリンクによって相互接続された計算を表現し、前記データフローグラフの少なくとも一つが少なくとも一つの入力データセットからデータのフローを受取り、且つ、前記データフローグラフの少なくとも一つが少なくとも一つの出力データセットにデータのフローを提供する、データストレージシステム；
データフローグラフと関連したデータセットの一つ以上のセットを特定するマッパ（mapper）であって、所与のセット内の各データセットが単一のデータセットの異なるバージョンを特定するための一つ以上の基準に一致する、マッパ；及び、
所与のセットにおける少なくとも二つのデータセットの間のマッピングを受け取るユーザインターフェースであって、データを前記マッピングのデータセットに提供するかまたはデータを前記マッピングのデータセットから受け取る一つのデータフローグラフに関連して前記データストレージシステム内の前記マッピングを格納するユーザインターフェース；
を含む。
【０００５】
他の概要において、データストレージシステムに格納されたデータをマッピングするシ

ステムは、
ノードを含むデータフローグラフの仕様を処理する手段であって、前記ノードがデータのフローを表現するリンクによって相互接続された計算を表現し、前記データフローグラフの少なくとも一つが少なくとも一つの入力データセットからデータのフローを受取り、且つ、前記データフローグラフの少なくとも一つが少なくとも一つの出力データセットにデータのフローを提供する、手段；
データセットの一つ以上のセットを特定する手段であって、所与のセット内の各データセットが単一のデータセットの異なるバージョンを特定するための一つ以上の基準に一致する、手段；
所与のセットにおける少なくとも二つのデータセットの間のマッピングを受け取るユーザインターフェースを提供する手段；及び、
データを前記マッピングのデータセットに提供するかまたはデータを前記マッピングのデータセットから受け取る一つのデータフローグラフに関連して前記ユーザインターフェースを通して受け取られた前記マッピングを格納する手段；
を含む。
【０００６】
他の概要において、データストレージシステムに格納されたデータをマッピングするためのコンピュータプログラムは、コンピュータ読み出し可能媒体に格納され、且つ、コンピュータに、
ノードを含むデータフローグラフの仕様を処理させる命令であって、前記ノードがデータのフローを表現するリンクによって相互接続された計算を表現し、前記データフローグラフの少なくとも一つが少なくとも一つの入力データセットからデータのフローを受取り、且つ、前記データフローグラフの少なくとも一つが少なくとも一つの出力データセットにデータのフローを提供する、命令；
データセットの一つ以上のセットを特定させる命令であって、所与のセット内の各データセットが単一のデータセットの異なるバージョンを特定するための一つ以上の基準に一致する、命令；
所与のセットにおける少なくとも二つのデータセットの間のマッピングを受け取るユーザインターフェースを提供させる命令；及び、
データを前記マッピングのデータセットに提供するかまたはデータを前記マッピングのデータセットから受け取る一つのデータフローグラフに関連して前記ユーザインターフェースを通して受け取られた前記マッピングを格納させる命令；
を含む。
【０００７】
態様は一つ以上の以下の特徴を含むことができる。
【０００８】
前記セットは前記ユーザインターフェース上に提示される。
【０００９】
前記一つ以上の基準への一致の数量化に従って順位付けられた可能性のあるマッピングのリストは、前記ユーザインターフェースを通して提示される。
【００１０】
前記可能性のあるマッピングのリストは、前記リスト内にてより高位に順序付けられた所与のデータセットのインスタンスである可能性がより高い候補を含む。
【００１１】
前記基準の一つは、前記一つ以上のデータセットを特定するマッパに組込まれている。
【００１２】
前記基準の一つは前記ユーザインターフェースから受け取られる。
【００１３】
前記可能性のあるマッピングの少なくとも一つは一つのデータセットを表現するデータフローグラフの構成要素を示し、且つ、前記可能性のあるマッピングの少なくとも一つは

データセットを表現しないデータフローグラフの構成要素を示す。
【００１４】
複数の構成要素を含むデータフローグラフのサブグラフは、ひとつのデータセットを表現する。
【００１５】
前記サブグラフはデータ構成要素を含む。
【００１６】
前記サブグラフは実行可能な構成要素を含む。
【００１７】
前記データセットの一つ以上のセットを特定することは、所与のセット内の一つのデータセットが他のデータセットと共通の一つ以上の特徴を有するか否かを判別するためのヒューリスティックスを使用することを含む。
【００１８】
前記特徴は前記データセットの表現におけるバイト及びレコードの量を含む。
【００１９】
前記特徴は前記データセットの表現の名称を含む。
【００２０】
前記特徴は前記データセットの表現の生成日を含む。
【００２１】
前記特徴は前記データセットの表現のデータフォーマットを含む。
【００２２】
前記マッピングのデータセットの少なくとも一つはデータ管理システムに知られているデータセットのグループに属する。
【００２３】
所与のセットのデータセット間のフォーマットマッピングが提供される。
【００２４】
前記マッピングは、前記データセットを追跡する前記データ管理システムにおけるレコードを指し示す識別子を含む。
【００２５】
データセットの変化に基づいて前記マッピングが更新される。
【００２６】
本発明の態様は以下の一つ以上の利点を有することができる。
【００２７】
バージョンを特定（識別）する基準に従ってデータセットの集合（セット）を特定することによって、一つのデータセットの二つのインスタンスの間の一致が純粋な手作業よりもより効率的になされ得る。さらに、少なくとも二つのデータセット間のマッピングを受取るユーザインターフェースを提供することにより、そのマッピングは、システムが純粋に自動化された場合よりもより正確になる。
【００２８】
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の請求項から明らかになる。
【図面の簡単な説明】
【００２９】
【図１】データフローグラフである。
【図２】データセットのマッパ及び関連する構成要素の概観図である。
【図３Ａ】データセットのマッパによって取扱われるシナリオの線図である。
【図３Ｂ】データセットマッパによって取扱われる異なるシナリオの線図である。
【図３Ｃ】データセットマッパによって取扱われる異なるシナリオの線図である。
【図３Ｄ】データセットマッパによって取扱われる異なるシナリオの線図である。
【図３Ｅ】データセットマッパによって取扱われる異なるシナリオの線図である。

【図４】データセットのマッパの動作フローチャートである。
【図５】データセットリンケージマッピングである。
【図６】データセットフォーマットマッピングである。
【発明を実施するための形態】
【００３０】
概観
データ処理要素はグラフの形式とすることができる。グラフに基づく計算は、構成要素（格納されたデータに対応するデータストレージ構成要素、または、実行可能な処理に対応する計算の構成要素の何れか）を表現するグラフにおける頂点（vertices）を有する有向グラフにより表現される「データフローグラフ」であってそのグラフにおける有向リンクまたは「辺（edges）」が構成要素間のデータのフロー（流れ）を表現する「データフローグラフ」を用いて実行される。データフローグラフ（これは、また、単に「グラフ」と呼ばれる。）はモジュールの統一体（modular entity）である。各グラフは、一つ以上の他の複数のグラフから構成されることができ、また、特定のグラフはより大きなグラフにおける一つの構成要素であることができる。グラフ開発環境（ＧＤＥ）は、実行可能なグラフを特定するとともにそのグラフの構成要素のパラメータを定義するユーザインターフェースを提供する。
【００３１】
図１を参照すると、データフローグラフ１０１の一例は、データフローグラフ１０１の実行可能構成要素１０４ａ－１０４ｊによって処理されるべきデータの集まりを提供する入力構成要素１０２を含む。たとえば、データセット１０２は、データベースシステムに関連するデータレコードあるいはトランザクション処理システムに関連するトランザクションを含むことができる。実行可能な構成要素の各々は、全体のデータフローグラフ１０１によって規定される計算の一部に関連している。作業要素（たとえばデータの集まりからの個別的なデータレコード）は一つの構成要素の一つ以上の入力ポートに入り、また、出力作業要素（幾つかの場合には、入力作業要素または入力作業要素の処理されたバージョン）は一般にその構成要素の一つ以上の出力ポートから出る。グラフ１０１において、構成要素１０４ｅ、１０４ｇ及び１０４ｊからの出力作業要素は、出力データ構成要素１０２ａ－１０２ｃ内に格納される。
【００３２】
データセットは、特定のデータの集まりを表わす（たとえばオブジェクト指向データベース内に格納された）オブジェクトである。データフローグラフのシステムのコンテクストにおいて、一つの構成要素は一つのデータセットを表わすことができる。このような場合、グラフはデータセットを表現する構成要素（即ち、単に「データセット構成要素」）と一つ以上の方法で相互に作用することができる。一つのデータセット構成要素は、所与のデータセットによって表現された物理的データをアクセスするための命令を含み、従って、一つのグラフはデータセットからの入力をデータセット構成要素を用いて受け取り、出力をデータセットにデータセット構成要素を用いて提供し、且つ、中間ステップにてデータセットのデータをデータセット構成要素を用いて処理することができる。データセット構成要素は、データセットオブジェクトの一つのインスタンスを含む所与のデータセットオブジェクトに関連する種々の情報を含むことができる。このようなシステムは、数十、数百、あるいは数千のグラフ及びこれに関連するデータセット構成要素を有することになる。このようなシステムの複雑度が増大するにつれて、種々のグラフとデータセット構成要素との間の関係を管理することは益々困難となる。システムにおける二つ以上のデータセット構成要素は同一のデータソースを表現することができ、また、そのようなデータセット構成要素は、各々、別のグラフ、グラフサブセットまたは実行可能な構成要素に関連付けることができる。
【００３３】
たとえば、一つの可能なシナリオにおいては、単一のデータセットはデータ管理システムに関連した二つ以上の場所に格納され得る。このシナリオにおいては、二つ以上のデー

タソースが、同一データの類似バージョンまたは同一バージョンを含む。このシステムにおける二つのグラフはこの単一のデータセットを取り扱うことになるが、各グラフは、別のデータベーステーブルまたは他の形式のデータセット構成要素を、別のデータファイルから読み取り且つ別のデータファイルへと書き込む。

【００３４】

類似したシナリオにおいて、所与のデータセットによって表現されたデータ（たとえばデータファイル）は二つ以上の場所に格納されるだけでなく、異なるデータストレージフォーマットを用いて解釈され得る。上述の例と同様、二つのグラフは、（フォーマットのみ異なる）同一のデータを含む二つの別のデータファイルに作用することができる。各データファイルは、同一のデータのインスタンスを含んでいるにもかかわらず、異なるデータ形式の構成（配列、arrangement）を有することがある。

【００３５】

別のシナリオにおいては、一つのグラフはデータセットのインスタンスを含む一つのデータファイルに作用し、また、他のグラフはそのデータセットのインスタンスを含むデータベーステーブルに作用する。このような場合、データファイル及びデータベーステーブルは一般に二つの異なるデータフォーマットを有するであろう。

【００３６】

他のシナリオにおいては、データ管理システムは同一のデータセットの異なるバージョンを各々異なる方法でアクセスできる。一つのグラフは、たとえばデータファイルを標準ファイル入出力メカニズムを用いて読み込むことによってデータセットのインスタンスを直接的にアクセスできる。他のグラフは、ネットワークを介して利用可能なデータ収納庫（repository）のような外部ソースに問い合わせることによってファイルを取得（retrieve）できる。また、あるグラフは、類似した外部問い合わせ（たとえばネットワークされたデータベースに対する問い合わせ）によってデータベーステーブルをアクセスできる。

【００３７】

また、データ管理システムは、同一データセットの異なるインスタンスを各々異なる方法で参照することができる。たとえば一つのグラフはパラメータに従って異なるデータ場所をアクセスできる。このようなパラメータは時間と共に多くのデータ場所を指し示すことができる。そのパラメータがグラフの実行間で変化すれば、複数回動作するグラフは異なる機会に異なる場所をアクセスできる。

【００３８】

あるシナリオにおいては、一つのグラフ内における一つのデータセットの表現は、単一構成要素ではなく、むしろ構成要素の集まり、たとえば、それ自体が複数の構成要素を有する一つのグラフとして実行されるグラフ内の「サブグラフ」である。この集まりは、一つ以上のデータセットを含むことができ、且つ、一つ以上の実行可能な構成要素を有することができる。

【００３９】

上述のシナリオのすべては、データ管理システムによって取り扱われるデータを視覚化し且つ解析することに対する問題を潜在的に提示する。ユーザが所与の一つのデータセットと相互に作用する構成要素の統合されたビューを必要とするとき、存在し得るデータセットの異なるインスタンスを調整（reconcile）するために種々のアプローチが用いられ得る。

【００４０】

一つのアプローチは、同一のデータセットの複数のインスタンスを特定し且つそれらの間にリンケージ（つながり、linkage）を生成する自動的なメカニズムである。しかしながら、幾つかの自動的なメカニズムは欠点、つまり以下に示す３つの欠点を有する。第１に、このメカニズムは、一つのデータセットの各インスタンスが特定な方法で格納されること、たとえば統一名称方式（unified
naming scheme）及びディレクトリ構造の下で格納されることを必要とする。これは、データ管理システムに関連するストレージシステム内において各々を特定し且つ場所を突き

とめる方法をそのメカニズムに提供する。しかしながら、この構成はデータ管理システムの柔軟性を限定すると共に、このシステムの幾つかの使用方法について過度に制限することになり得る。

【００４１】

第２に、動作についての幾つかのシナリオにおいて、そのメカニズムは同一のデータセットのインスタンスを正しく特定せず、また、正しいリンケージを形成しない場合がある。たとえば、このことは、データセットを外部参照エンティティを用いてアクセスし、且つ、その自動的なメカニズムがそのエンティティに対するアクセスを有しない場合に発生する可能性が高い。同様に、このことは、構成要素がデータセットをパラメータリストにおける独立パラメータに従ってアクセスし、且つ、この自動的なメカニズムがそのパラメータリストをアクセスする方法または解釈する方法を有しない場合に発生する可能性が高い。さらに、このことは、一つのデータセットが、一つ以上のデータセット構成要素及び実行可能な構成要素（たとえばサブグラフ）から構成される複雑なエンティティによって表現されているときに発生する可能性が高い。自動的なメカニズムが、構成要素のどの特定な組み合わせが特定のデータセットを表現するかを識別できない場合もある。

【００４２】

第３に、そのメカニズムは、データセットのインスタンスの間に冗長または不要のリンケージを形成する場合がある。たとえば、データ管理システムによって取り扱われるデータセットのいくつかは、たとえばエラーログのような外部からのデータを表現する場合がある。これらのデータセットの間のいかなるリンケージも不要である。さらに、データ管理システムによって取り扱われるデータセットのインスタンスのいくつかは、冗長な（重複する）インスタンス（たとえばキャッシュされたデータまたはデータの他の一時的なコピー）である場合がある。この形式のデータに接続するリンケージは直ちに陳腐化し、且つ、データ管理システムを分析するユーザを当惑させるであろう。

【００４３】

代替のアプローチは、ユーザが同一データセットのインスタンスをユーザインターフェースを介して手動で統合するシステムである。ユーザが一つのデータセットの複数のインスタンスの間の本質的なリンケージを見逃す可能性は低く、且つ、ユーザが一つのデータセットの複数のインスタンスの間の冗長または不要なリンケージを生成する可能性も低い。しかしながら、データ管理システムが数百・数千の構成要素を有する場合、ユーザが手動で必要なリンケージを生成するために必要な時間はとてつもなく長い。

【００４４】

部分的に自動化されたアプローチにおいては、データセットマッパ（dataset mapper）が使用されることにより、幾つかの自動的な分析を提供され、且つ、大きな及び／または複雑なシステムのユーザにとって手が出せないことがない方法でそのユーザとの所定の交流を可能とする。

【００４５】

図２は、例示的なデータセットマッパ１００の一実施形態のブロック線図であって、関連する主要な要素間の相互関係を示す。データセットマッパ１００は、一つ以上のグラフ１８０、１８０ａ、１８０ｂ、１８０ｃのセットを解析することができる。各グラフは、一つ以上のデータセット構成要素１８２、１８２ａ、１８２ｂに関連し、各データセット構成要素は、データファイル、データベーステーブル、サブグラフまたはデータセットを表現する他の種類の構成要素に対応する可能性がある。マッパ１００は、同一のデータセット１８６のインスタンスを含むデータセットの構成要素間のリンケージを形成する目的でグラフを解析する。マッパ１００は、組込み規則１１０、ユーザ定義規則１２０及びヒューリスティックス（経験則、heuristics）１３０の組合せに従って各データセット構成要素を処理してデータセット構成要素１８２がデータ管理システム１７０に知られたデータソース１７６、１７６ａ、１７６ｂを表現する幾つかのデータセットの一つのインスタンスを含むか否かを判別する。マッパ１００はこの情報をユーザインターフェース１６０に渡し、インターフェース１６０はデータセット構成要素１８２に対応する正しいデータ

セットがあればこの正しいデータセットをユーザ１６２に選択させることを許容する。たとえば、ユーザインターフェース１６０は、単一のデータセットの異なるバージョンまたはインスタンスを特定する一つ以上の基準への合致に基づいて可能性のある候補のマッピングのリストを提示する。組込み規則、ユーザ定義規則及びヒューリスティックスに基づく基準を含むそのような基準の例について以下に詳述する。候補マッピングのリストは一つ以上の基準への合致の数量化に従って順位付け（order）できる（たとえば、所与のデータセットのインスタンスである可能性がより大きい候補はリストにおいてより高位に順序付けられる）。次に、マッパ１００は、データセット構成要素１８２がデータソース１７６を表現するデータセットのインスタンスを含むことを示すデータセットリンケージマッピング１４０を発生する。

【００４６】

さらに、データセット構成要素１８２は、対応する連結されたデータソース１７６のフォーマット１７４とは異なるデータフォーマット１８４を有することができる。データ管理システム１７０の要求に依り、ユーザはデータセットのすべてのインスタンスに対して単一のデータフォーマットを確立することを選択することができる。このシステムは各データソース１７６、１７６ａ、１７６ｂに対してフォーマット１７４、１７４ａ、１７４ｂを格納する。代替として、ユーザは、データセット構成要素１８２のフォーマット１８４と、対応するデータソース１７６の確立されたフォーマット１７４と、の間に随意（optional）のマッピング１４２を生成することを選択することができる。その随意のマッピング１４２は、データ管理システム１７０がデータセットの各インスタンスに対するデータの形式についての情報を保持することを許容する。

【００４７】

また、マッパ１００は、ユーザが、実行可能な構成要素と、他のリンケージは有しないであろう単一のデータセット構成要素と、の間のリンケージを示すことを可能とする。たとえば、データセット構成要素は、単一の読取者を伴うソースデータセット、または、単一の書込者を伴う目標データセットに対応することができる。データセットオブジェクトが、システムに存在し、且つ、正しいレコードフォーマット、文書、データプロフィール等の他の関連するメタデータを有すれば、リンケージはデータセット構成要素を正しいデータセットにマップさせることができる。

【００４８】

マッピング処理

マッパ１００は、複雑なデータ管理システムに生ずる共通のシナリオを取扱うことができる。第１のシナリオにおいては、図３Ａに示したように、一つのグラフ２１０は出力としてデータセット構成要素２１２を提供し、他のグラフ２２０は入力として異なるデータセット構成要素２２２を受取る。各データセット構成要素は、同一のデータセット２１６のインスタンスを含む。このデータセットは、データ管理システムに知られたデータソース１７６を表現するデータセットと同じであり得る。さらに、第１のデータセット構成要素２１２は第２のデータセット構成要素２２２に属するフォーマットと同じデータフォーマット２１４を有し、あるいは、代替として、第２のデータセット構成要素は異なるフォーマット２２４を有してもよい。マッパ１００は、第２のデータセット構成要素２２２を第１のデータセット構成要素２１２によって表現されたデータセット２１６のインスタンスであるとして特定し、適切なリンケージマッピング１４０を生成することができる。

【００４９】

第２のシナリオにおいては、図３Ｂに示したように、グラフ２３０は、外部ソース２３９に対する外部参照２３８を用いて外部データセット構成要素２３２に関連付けられる。その外部データセット構成要素２３２は、データフォーマット２３４を有し、且つ、データセット２３６のインスタンスである。第１のシナリオと同様、外部データセット構成要素によって表されたデータセット２３６は、データ管理システム１７０に知られたデータソース１７６を表現するデータセットであり得る。マッパ１００は、この外部データセット構成要素２３２を他のデータセットのインスタンスであるとして特定し、適切なリンケ

ージマッピング１４０を生成することができる。
【００５０】
第３のシナリオにおいては、図３Ｃに示したように、グラフ２４０は、パラメータリスト２４７内のパラメータ２４８を用いてデータセット構成要素２４２に関連付けられる。参照されるデータセット構成要素２４２は、データフォーマット２４４を有し、且つ、データセット２４６のインスタンスである。第１、第２のシナリオと同様、参照されるデータセット構成要素によって表されたデータセット２４６は、データ管理システム１７０に知られたデータソース１７６を表現するデータセットであり得る。マッパ１００は、この参照されるデータセット構成要素２４２を他のデータセットのインスタンスであるとして特定し、適切なリンケージマッピング１４０を生成することができる。
【００５１】
第４のシナリオにおいては、図３Ｄに示したように、グラフ２５０は外部ソース２５９に対する外部参照２５８を用いて外部構成要素２５１に関連付けられる。外部構成要素２５１は、データセット構成要素ではなく、むしろ実行可能な構成要素のような他の種類の構成要素である。マッパ１００は、この外部構成要素２５１を、データセットリンケージマッピング処理に適用できないものとして特定することができる。
【００５２】
第５のシナリオにおいては、図３Ｅに示したように、グラフ２６０は、それ自体複数の構成要素から構成されるサブグラフ構成要素２６３に関連している。これらの構成要素は、少なくとも一つのデータセット構成要素２６２を含み、且つ、この例では一つ以上の実行可能な構成要素２６１ａ、２６１ｂ、２６１ｃを含む。このシナリオにおいては、単一エンティティとしてのサブグラフ２６３は、少なくとも一つのデータセットを表現する。他のサブグラフの例は複数のデータセット構成要素及びゼロを含む多数の実行可能な構成要素を含むことができる。さらに、サブグラフ２６３は複数の出力２６５ａ、２６５ｂを有する。各出力は、データセットの異なるインスタンスを、その出力を受け取る構成要素に提供することができる。また、サブグラフの他の例は、多数の入力を有することもできる。さらに別のサブグラフの例は、それぞれのデータセットに対応する入力または出力を有しない場合もある。サブグラフが少なくとも一つのデータセットを表現している場合には、マッパ１００は、サブグラフ２６３を少なくとも一つのデータセットのインスタンスであるとして特定し、少なくとも一つの適切なリンケージマッピング１４０を生成することができる。
【００５３】
マッパの動作シーケンスの例を図４に示す。ステップ３０２において、マッパは第１に一つのグラフに関連付けられた要素のうちのどの要素がデータセットを表現するかを特定する。一般に、グラフは一つ以上の入力及び出力を有し、また、各入力及び各出力はデータセットのインスタンスであろう。また、各グラフは、ある中間ステップにおいて、データセットのインスタンスを取扱うことができる。この結果、各グラフは、データセットの候補となり得る複数の構成要素に接続され得る。幾つかの場合には、データ管理システムは、ある構成要素がデータセットを表現するか否かについての情報を含む「構成要素の特性」についての情報を有する。このような場合、ステップ３０４において、マッパは可能性のあるデータセット構成要素をデータセットの候補テーブルに加える。幾つかの場合、構成要素は、データセット構成要素及び実行可能な構成要素を含む複数の構成要素から構成されたサブグラフであり得る。サブグラフは一つのデータセットの少なくとも一つのインスタンスを表現することができるであろう。従って、マッパは、そのようなサブグラフの総てのリストを編集（コンパイル）し、ステップ３０４の一部としてそれらをデータセットの候補テーブルに加える。他の幾つかの場合において、構成要素の特質（nature）がデータ管理システムに利用され得ないものであることがある。その構成要素は、外部エンティティへの参照を介してアクセスされ得る。ここで、その外部エンティティへの参照とは、データベーステーブル、インタネットサーバを指し示すユニフォームリソースロケイタ、パラメータリスト内のパラメータ、または、他の形式の参照に対する問合せ（クエリ

ー）である。これらの場合、一般に、マッパは、その（外部エンティティへの）参照によって指し示されたエンティティを独立にアクセスするための手段を有しない。従って、マッパは、そのような参照の総てのリストを編集（コンパイル）し、ステップ３０４の一部としてそれらをデータセット候補テーブルに加える。

【００５４】

次に、ステップ３０６において、所与のデータセットの候補に対し、マッパは、データセットの候補がマップする可能性のある既知のデータセットのリストを発生する。マッパは、ユーザ定義規則、組込み規則及びヒューリスティックスの組合せを用いて、既知のデータセットの何れがデータセットに候補にマップされるかを評価する。

【００５５】

次に、ステップ３０８において、ユーザは、そのデータセットの候補に対応する既知のデータセットを選択する。また、ユーザは、提案された既知のデータセットの何れもが正しい合致でない場合、すべての既知のデータセットのリスト全体をアクセスするかもしれない。加えて、ユーザは、データセットの候補がデータセットでないことを示すことができる。たとえば、遠隔サーバへの参照は、遠隔実行可能手続（データエンティティでない）への要求であろう。他の例として、データセットの候補はデータを表現できるが、そのデータは、エラーログのようなデータ管理システムに関係しない種類のデータである場合がある。この場合、ユーザは、ユーザインターフェースに対してこのデータはマッピング処理において無視すべきであることを指示することができる。

【００５６】

次に、ステップ３１０において、ユーザは新しくマップされたデータセットのデータフォーマットを特定する。システムは、データフォーマットテンプレートの集合を有し、その一つが選択され得る。代替として、ユーザはユーザインターフェース内に新しいデータフォーマットを生成することができる。

【００５７】

次に、ステップ３１２において、マッパは、この情報を用いてデータセットの候補に対するリンケージマッピング、及び、随意に、フォーマットマッピングを発生する。

【００５８】

次に、マッパがすべてのデータセットの候補を処理していなければ、マッパは、ステップ３０８、３１０及び３１２の繰り返しにおけるリンケージ発生のために、次のデータセットの候補をユーザに提供する。

【００５９】

次に、ステップ３１４において、ユーザはデータ管理システムに関連した構成要素を見て、グラフとデータセット構成要素との間の関連の視覚化が正確であることを構成要素間の新しいリンケージに基づいて保証する。ステップ３１６において、ユーザはリンケージ及びフォーマットマッピングに対して調整するオプションを有する。

【００６０】

最後に、ステップ３１８において、マッパは、リンケージ及びフォーマットマッピングをデータ管理システムに引き渡す。そのマッピングは、一つ以上のグラフと一緒に格納され、またはデータ管理システムに関連する別のストレージエンティティ内に格納され、もしくは、他の手段によって格納されることができる。

【００６１】

データセットマッピングのメンテナンス

マッパ１００は、データセットリンケージの完全性（インテグリティ、integrity）に影響する複数のシナリオを取り扱うことができる。

【００６２】

第１のシナリオは、新構成要素がデータ管理システム１７０に追加されたときに新しいデータセットの候補を特定することを含む。このシナリオの下では、マッパ１００は各構成要素を分析し且つ可能性のあるリンケージをユーザに提示する。マッパ１００は、どのような新しい構成要素にも作用して適切なリンケージを必要に応じて発生する。

【００６３】
第２のシナリオは、データ管理システム１７０が時間と共に変化したときに、現存のリンケージをメンテナンスすることを含む。たとえば、データ管理システムが関連するグラフの通常の動作中にデータセットの新インスタンスが出現することがある。他の一つの例として、データセットがそのアイデンティティ、たとえばデータ管理システムにおけるその名称及び場所、を変更する場合がある。さらなる例として、データセット全体が削除されている場合もある。さらなる一つの例として、データセット候補は、リンケージ生成の先の回において見落とされていたかも知れず、その結果、リンケージの集まりが不完全な場合がある。マッピングシステムのユーザインターフェース１６０は、ユーザ１６２に、現存のリンケージを修正させて不完全または期限切れのマッピングを修復させることを許容する。

【００６４】
第３のシナリオは、既知のパターンに判で押したように従うデータセットの参照のリンケージを自動的に更新することを含む。たとえば、グラフは、パラメータリスト２４７において参照されるデータセットを取り扱うことができる。このようなパラメータリストは時間と共に変化する。パラメータリストがデータ管理システムに知られた標準フォーマットに従うのであれば、マッパはパラメータリスト内の変化を特定し且つ現存のリンケージをそれに応じて更新することができる。

【００６５】
データセットリンケージのマッピング
図５に示したように、データセットリンケージマッピング１４０は、構成要素名４０２、データセット名４０４、データセット形式４０６、フォーマット４０８、マスタデータセット場所４１０及びフラグ４１２を含む。構成要素名４０２はデータセット構成要素またはそのデータセットのインスタンスを表現するサブグラフである。データセット名４０４はこの構成要素によって表現されたデータセットを指し示す識別子である。データセット形式４０６は、データファイル、データベーステーブルまたは他の形式等のこのデータセットのインスタンスが分類されるカテゴリを示す。フォーマット４０８は、このデータセットのインスタンスがそのデータを表現するのに用いるフォーマットまたは配列である。マスタデータセット場所４１０は、このデータセットを追跡するデータ管理システム内のレコードを指し示す識別子である。最後に、フラグ４１２は、たとえばユーザがこのデータセットのインスタンスをデータ管理システムに適用できないものとして特定し、このインスタンスはリンケージの集合から排除されるべきであると特定した場合に、このデータセットのインスタンスは無視されるべきか否かを示す。

【００６６】
組込み規則
マッパ１００は、データ管理システムの標準的な慣例（standard conventions）に従って動作する組込み規則１００の集合を有する。データセット構成要素が組込み規則１１０に従う場合、マッパはデータセット構成要素に対応するデータセットを最高位精度で特定することができる。規則の一つの実施態様において、データセットの候補を含む外部参照データベーステーブルは、データ管理システムによって用いられる標準化されたディレクトリ構造下で永続記憶装置内に配置されなければならない。さらに、パラメータに従って外部参照データセット構成要素をアクセスするグラフは、データ管理システムがまたアクセスし且つ分解できるパラメータを用いなければならない。さらに、データセット構成要素のフォーマットは、永続記憶装置に内にて利用可能であり、且つ、データ管理システムによってアクセスできるものでなければならない。データ管理システムに依り、他の組込み規則でもよい。

【００６７】
ユーザ定義規則
データセットの候補を特定するためにマッパが使用する組込み規則に加えて、マッパ１００は、随意のユーザ定義規則１２０の集まりを有する。これらのユーザ定義規則１２０

は、どれがユーザの特定のデータ管理システムに適用可能であるかに応じて、ユーザによって使用可能または使用不能に設定され得る。一実施態様において、マッパは六個の選択が自由なユーザ定義規則を有する。データベーステーブルという名目の一部の情報（たとえば、そのテーブルを定義したユーザについての情報等）がテーブルのアイデンティティを不明瞭にする場合、マッパはデータベーステーブルという名目の一部の情報を無視することができる。さらに、マッパは、データベーステーブルの名称からこの情報を除去することができる。さらに、マッパは、データ管理システムに関連するデータセットに関係のないデータを含むことが知られている特定のカテゴリのデータファイルを、無視することができる。そのようなカテゴリは、データファイル形式またはデータファイルエクステンションであろう。さらに、マッパは、参照をパラメータリスト内の特定のパラメータへと変換し、且つ、この参照をパラメータ自身の名称に置換できる。さらに、マッパは、パラメータに対する参照を完全に除去することができる。ユーザは、また、マッパが従う他の規則を生成することもできる。

【００６８】

ヒューリスティックス

データセット候補を評価するための組込み規則及びユーザ定義規則に続いて、さらにマッパ１００はヒューリスティックス１３０の集合を用いる。ヒューリスティックス１３０は、マッパに、所与のデータセット構成要素の特徴を解析させることを許容し、これらの特徴を既知のデータセットと比較させる。既知のデータセットに類似した特徴を有するデータセット構成要素はそのデータセットの一つのインスタンスである可能性が高い。一つの実施態様においては、マッパは二つのヒューリスティックスを用いる。一つのヒューリスティックスは所与のデータセット構成要素のデータの特徴である。たとえば、データセット構成要素に関連するデータが、既知のデータセットに関連するデータと同一量のバイト及びレコードを有するときには、そのデータセット構成要素はそのデータセットのインスタンスの可能性が高い。さらに、データセット構成要素が、既知のデータセットと類似の名前または日付のクリエーションを有するならば、そのデータセット構成要素はそのデータセットのインスタンスである可能性が高い。第２のヒューリスティックスはデータセット構成要素のデータフォーマットである。データセット構成要素が既知のデータセットとデータフォーマットを共有していれば、そのデータセット構成要素はそのデータセットのインスタンスの可能性が高い。このヒューリスティックスは、複数の別個のデータセットが同一のデータフォーマットを用いている状況では、信頼性が低い。

【００６９】

データセットのフォーマット及びマッピング

データソースを表現する各データセットは、データセットの各要素に対して、その要素がどのデータ形式を表しているかを示す関連データフォーマットを有する。たとえば、データベーステーブルのデータフォーマットは、所与のレコード内の各フィールドのデータ形式を示す。データ管理システム１７０は、データソース１７６、１７６ａ、１７６ｂを表現する各データセットに対する単一のデータフォーマット１７４、１７４ａ、１７４ｂを保持する。

【００７０】

マッパ１００が新たなデータセット１８６を表現するデータセット構成要素１８２に遭遇した場合、マッパ１００は、データセット構成要素１８２のデータフォーマット１８４に基づいてデータ管理システムによって格納されるべき対応するデータフォーマットを生成する。

【００７１】

データセット構成要素１８２がデータソース１７６を表現する既知のデータセットを表している幾つかの場合には、データセット構成要素１８２は、データソース１７６を表現する既知のデータセットのデータフォーマット１７４とは異なるデータフォーマット１８４を有する。データ管理システム１７０は、データソース１７６を表すデータセットを、存在するかもしれないデータセットのインスタンスの数とは関係なく、単一のエンティテ

ィとして取扱う。その結果、このような状況が発生したときには、データ管理システム１７０はマッパ１００を頼りにして異なるフォーマット１７４、１８４を統合させる。一実施態様においては、マッパは、ユーザ及びデータ管理システムの要求に依存して四つの異なる方法のうちの一つの方法にて各状況に対処することができる。ユーザ１６２は、各状況に対する統合のための四つの方法のうちのどれでも一つを選択することができる。

【００７２】

統合のための第１の方法の下では、マッパ１００は、データセット構成要素１８２のデータフォーマット１８４を、そのデータセットのマスタデータフォーマットとして用い、それに応じてデータ管理システム１７０を更新する。

【００７３】

統合のための第２の方法の下では、マッパ１００は、現存のデータセットのデータフォーマット１７４を、そのデータセットのマスタデータフォーマットとして用い、それに応じてデータ管理システム１７０を更新する。

【００７４】

統合のための第３の方法の下では、マッパ１００は、両方のデータフォーマットを保持し、各データフォーマットのフィールド間のマッピング１４２を発生する。図６に示したように、データフォーマットマッピング１４２は、データセットフォーマット５１０のフィールド５１２ａ、５１２ｂ、５１２ｃのどれがデータセットインスタンス（たとえば、データセットの構成要素）のフォーマットのフィールド５２２ａ、５２２ｂ、５２２ｃのどれに対応するかを示す。

【００７５】

統合のための第４の方法の下では、マッパは、いずれか一方のデータフォーマットとして作用できる新しい結合データフォーマットを発生する。

【００７６】

一般的なコンピュータの実施形態

上述したデータセットマッピングのアプローチは、コンピュータ上で実行されるソフトウェアを用いて実現できる。たとえば、そのソフトウェアは、一つ以上のプログラムされたまたはプログラム可能なコンピュータシステム（分散型、クライアント／サーバ型、または、グリッド型等の種々アーキテクチャ型である）上で実行される一つ以上のコンピュータプログラムにおける手続を形成する。その各コンピュータシステムは、少なくとも一つのプロセッサ、少なくとも一つの記憶システム（揮発性メモリ、及び不揮発性メモリ、及びまたは記憶素子を含む。）、少なくとも一つの入力装置またはポート、及び、少なくとも一つの出力装置またはポートを含む。そのソフトウェアは、たとえば、データフローグラフの設計及び構成に関係した他のサービスを提供するより大きなプログラムの１つ以上のモジュールを形成できる。グラフのノード及び要素は、コンピュータ読出し可能媒体に格納されたデータ構造、または、データ貯蔵庫（保管所）に格納されたデータモデルに合致する他の系統的なデータとして実現され得る。

【００７７】

そのソフトウェアは、汎用または専用のプログラマブルコンピュータによって読み出し可能なＣＤ－ＲＯＭ等の記憶媒体上に提供され、或いは、そのソフトウェアが実行されるコンピュータにネットワークの通信媒体を用いて（伝播信号に符号化されて）配信される。すべての機能は、専用のコンピュータ上で、または、コプロセッサのような専用ハードウェアを用いて実行され得る。ソフトウェアは、ソフトウェアによって特定される異なる計算の部分が異なるコンピュータによって実行されるという分散方式によって実行され得る。好ましくは、上述の各コンピュータプログラムは、汎用または専用プログラマブルコンピュータによって読み出し可能な記憶媒体またはデバイス（たとえば、固体メモリまたは媒体、もしくは磁気的または光学的媒体）に格納され或いはダウンロードされ、記憶媒体またはデバイスがコンピュータシステムによって読み出されて上述した手順を実行するときに、コンピュータを構築及び動作させる。また、本発明のシステムは、コンピュータプログラムとともに構築されたコンピュータ読み出し可能記憶媒体として実現されること

が考えられる。ここで、そのように構築された記憶媒体は、上述の機能を実行するために特別且つ予め定義された様式でコンピュータシステムを作動させる。

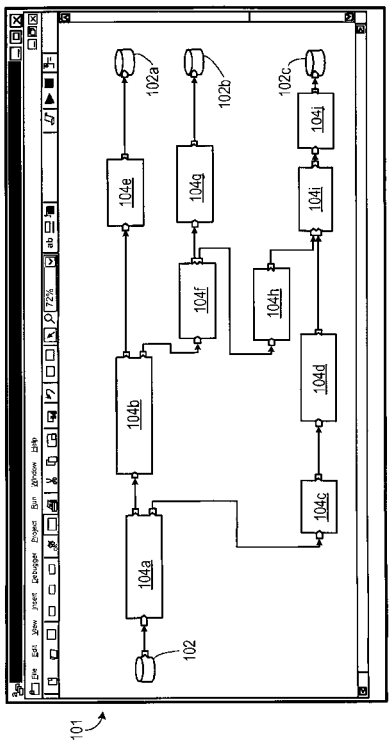
【００７８】

多くの本発明の実施例を記載した。しかしながら、本発明の精神及び範囲から逸脱することなく多種の変更がなされ得ることが理解されよう。たとえば上述のいくつかのステップは、順序が独立であり、従って上述の順序と異なる順序で実行され得る。

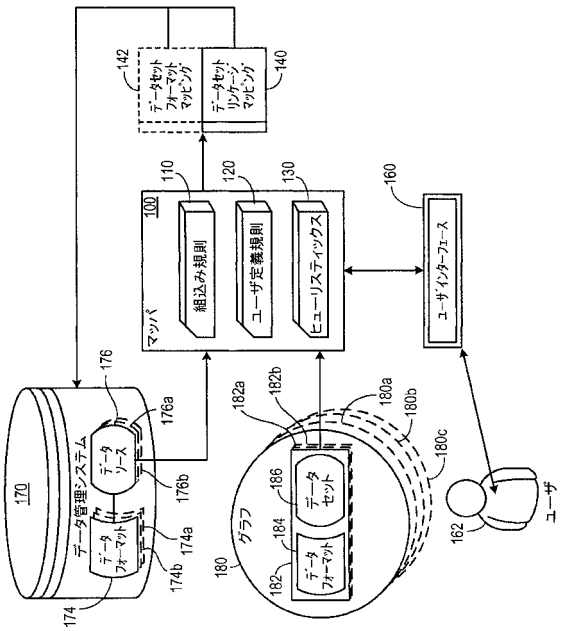
【００７９】

上述の記述は説明のためであり、添付された本発明の特許請求の範囲によって定義される発明の範囲を限定する意図はないことが理解されるべきである。たとえば、上述の機能ステップの多くは、全体の処理に実質的に影響を及ぼすことなく異なる順序で実行され得る。他の実施例は特許請求の範囲の請求項の範囲内にある。

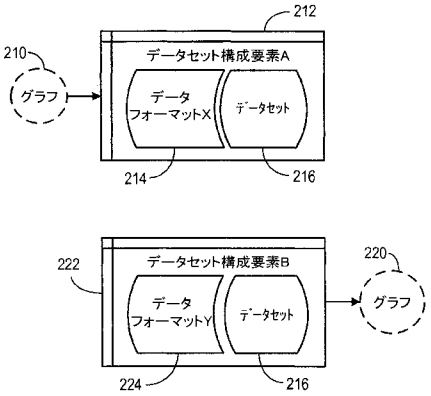
【図１】



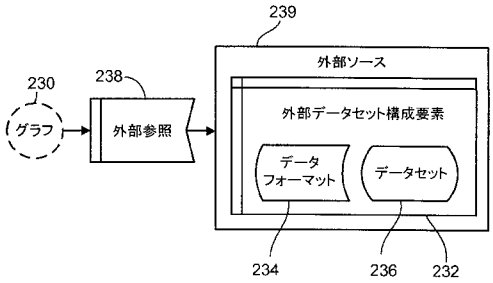
【図２】



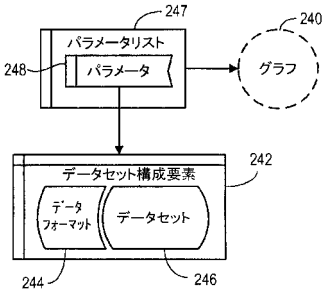
【図３Ａ】



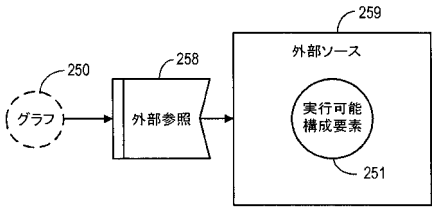
【図３Ｂ】



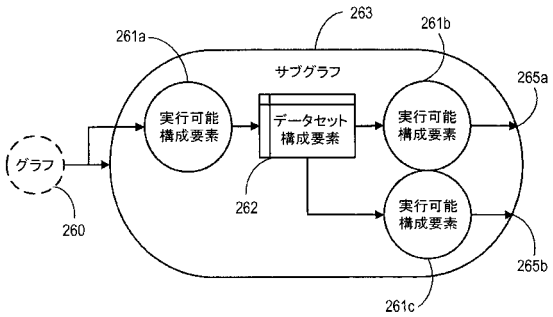
【図３Ｃ】



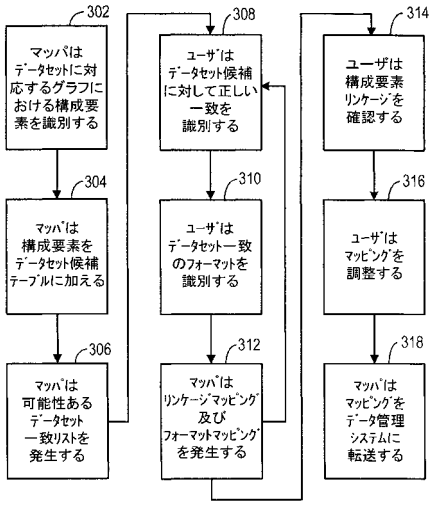
【図３Ｄ】



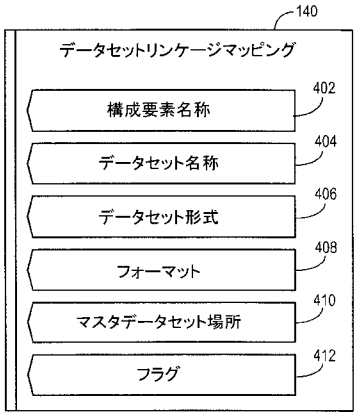
【図３Ｅ】



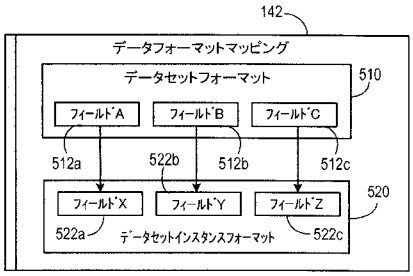
【図４】



【図５】


140
データセットリンケージマッピング
構成要素名称
402
データセット名称
404
データセット形式
406
フォーマット
408
マスタデータセット場所
410
フラグ
412

【図６】


142
データフォーマットマッピング
データセットフォーマット
510
フィールド'A
フィールド'B
フィールド'C
512a
522b
512b
512c
520
フィールド'X
フィールド'Y
フィールド'Z
522a
データセットインスタンスフォーマット
522c

フロントページの続き

(72)発明者　ワイス　アダム
アメリカ合衆国　０２４２０　マサチューセッツ州　レキシントン　ローソン　アベニュー　１５

審査官　岡本　俊威

(56)参考文献　特開平１１－１４３７５５（ＪＰ，Ａ）
米国特許第０７０８００８８（ＵＳ，Ｂ１）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｔ　　１／００
Ｇ０６Ｆ　１２／００