



(12) 发明专利

(10) 授权公告号 CN 102232212 B

(45) 授权公告日 2015. 11. 25

(21) 申请号 200980148087. 4

(22) 申请日 2009. 12. 01

(30) 优先权数据

61/119, 164 2008. 12. 02 US

(85) PCT国际申请进入国家阶段日

2011. 06. 01

(86) PCT国际申请的申请数据

PCT/US2009/066210 2009. 12. 01

(87) PCT国际申请的公布数据

W02010/065511 EN 2010. 06. 10

(73) 专利权人 起元技术有限责任公司

地址 美国马萨诸塞州

(72) 发明人 蒂姆·韦克林 亚当·韦斯

(74) 专利代理机构 北京林达刘知识产权代理事
务所(普通合伙) 11277

代理人 刘新宇

(51) Int. Cl.

G06F 7/00(2006. 01)

(56) 对比文件

US 7080088 B1, 2006. 07. 18, 说明书第 4 栏第 1-23 行, 第 5 栏第 52 行-第 6 栏第 18 行, 第 6 栏第 43-65 行, 第 6 栏第 57 行-第 7 栏第 25 行, 第 8 栏第 1-46 行, 第 11 栏第 40 行-第 12 栏第 32 行, 第 12 栏第 45 行-第 13 栏第 43 行, 第 13 栏第 60 行-第 14 栏第 21 行, 第 14 栏第 26-44 行, 第 17 栏第 54-67 行, 第 18 栏第 43 行-第 19 栏第 26 行, 第 19 栏第 30-62 行, 第 20 栏第 41-67 行, 第 21 栏第 20 行-第 22 栏第 3 行, 第 22 栏第 7-24, 38-65 行。

US 2007011208 A1, 2007. 01. 11, 说明书第 0011-0013, 0043-0045, 0063-0064, 0075-0078, 0080, 0083, 0085-0089 段。

US 2004056908 A1, 2004. 03. 25, 全文。

US 2001014890 A1, 2001. 08. 16, 全文。

US 2005187984 A1, 2005. 08. 25, 全文。

审查员 李敏

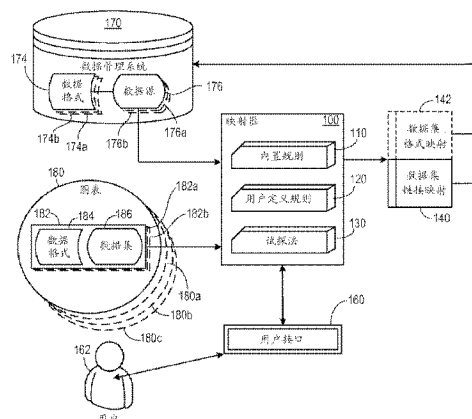
权利要求书4页 说明书9页 附图7页

(54) 发明名称

在数据管理系统中映射数据集的实例

(57) 摘要

一种用于映射存储在数据存储系统 (170) 的数据以由计算机系统使用的方法, 包括: 处理包括代表由表示数据的流的链路相互连接的计算的节点的数据流图表 (180) 的规范。至少一个数据流图表从至少一个输入数据集接收数据的流以及至少一个数据流图表向至少一个输出数据集提供数据的流。映射器 (100) 识别一个或多个数据集的集合。在给定集合中的每个数据集匹配用于识别单个数据集的不同版本的一个或多个准则。提供用户接口 (160) 来接收在给定集合的至少两个数据集之间的映射。与数据流图表关联来存储在用户接口上接收的映射, 该数据流图表向该映射的数据集提供数据或从该映射的数据集接收数据。



1. 一种用于映射存储在数据存储系统中的数据的方法,包括:

接收第一数据流图表的规范,其从至少一个输入数据集接收数据并且向至少一个输出数据集输出数据;

接收第二数据流图表的规范,其从至少一个输入数据集接收数据并且向至少一个输出数据集输出数据;

识别第一数据集候选者,所述第一数据集候选者的表示属于第一数据流图表,并且所述第一数据集候选者匹配用于识别第一数据集的不同版本的一个或多个准则;

识别第二数据集候选者,所述第二数据集候选者的表示属于第二数据流图表,并且所述第二数据集候选者匹配用于识别所述第一数据集的不同版本的一个或多个准则;

提供用户接口以接收指示所述第一数据集候选者和所述第二数据集候选者两者表示同一数据集的映射;以及

与数据流图表关联来存储由用户接口接收的所述映射,所述数据流图表向所述第一数据集候选者或者所述第二数据集候选者提供数据或从所述第一数据集候选者或者所述第二数据集候选者接收数据。

2. 根据权利要求1所述的方法,包括在用户接口上呈现所述第一数据集候选者和所述第二数据集候选者。

3. 根据权利要求1所述的方法,包括在用户接口上呈现数据集之间的映射的列表,所述映射根据对一个或多个准则的匹配的数量而排序。

4. 根据权利要求3所述的方法,其中映射的列表包括在列表中排序较高的给定数据集的版本的候选者。

5. 根据权利要求3所述的方法,其中准则之一内置于映射器中,该映射器识别所述第一数据集候选者和所述第二数据集候选者。

6. 根据权利要求3所述的方法,其中准则之一是从用户接口接收的。

7. 根据权利要求3所述的方法,其中至少一个映射指示表示所述数据集候选者之一的数据流图表的组件,并且至少一个映射指示不表示所述数据集候选者之一的数据流图表的组件。

8. 根据权利要求7所述的方法,其中包括多个组件的数据流图表的子图表表示所述数据集候选者之一。

9. 根据权利要求8所述的方法,其中该子图表包括数据组件。

10. 根据权利要求8所述的方法,其中该子图表包括可执行组件。

11. 根据权利要求1所述的方法,其中识别所述第一数据集候选者和所述第二数据集候选者包括使用试探法以确定所述第一数据集候选者是否具有与所述第二数据集候选者共同的一个或多个特征。

12. 根据权利要求11所述的方法,其中该特征包括在所述数据集候选者之一的表示中的字节和记录的数量。

13. 根据权利要求11所述的方法,其中该特征包括所述数据集候选者之一的表示的名称。

14. 根据权利要求11所述的方法,其中该特征包括所述数据集候选者之一的表示的创建的日期。

15. 根据权利要求 11 所述的方法,其中该特征包括所述数据集候选者之一的表示的数据格式。

16. 根据权利要求 1 所述的方法,其中所述第一数据集候选者和所述第二数据集候选者属于数据管理系统已知的数据集的组。

17. 根据权利要求 1 所述的方法,还包括提供所述第一数据集候选者和所述第二数据集候选者之间的格式映射。

18. 根据权利要求 1 所述的方法,其中还包括存储标识符,所述标识符指向保持所述第一数据集候选者和所述第二数据集候选者的踪迹的数据管理系统中的记录。

19. 根据权利要求 1 所述的方法,还包括基于在所述第一数据集候选者和所述第二数据集候选者的至少一个中的改变更新数据集之间所存储的映射。

20. 根据权利要求 1 所述的方法,其中所述第一数据集的每个版本与不同的数据流图表、数据流图表子集或可执行组件关联。

21. 根据权利要求 1 所述的方法,其中所述第一数据集的每个版本存储在与数据存储系统关联的不同位置。

22. 根据权利要求 1 所述的方法,其中所述第一数据集的每个版本使用不同的数据存储格式来解释。

23. 根据权利要求 1 所述的方法,其中使用在所述数据流图表的执行之间改变的参数来访问所述第一数据集的每个版本。

24. 根据权利要求 1 所述的方法,其中数据流图表中的至少一个包括代表由表示数据的流的链路相互连接的计算的节点,所述数据流图表从所述至少一个输入数据集接收数据的流并且向所述至少一个输出数据集提供数据的流。

25. 根据权利要求 1 所述的方法,其中所述第一数据流图表为数据流图表的子图表并且所述第二数据流图表为所述数据流图表的子图表。

26. 一种用于映射存储在数据存储系统中的数据的系统,该系统包括:

存储器,用于存储指令;

处理器,用于执行所存储的指令;

数据存储系统,其存储:

第一数据流图表的规范,其从至少一个输入数据集接收数据并且向至少一个输出数据集输出数据;以及

第二数据流图表的规范,其从至少一个输入数据集接收数据并且向至少一个输出数据集输出数据,

映射器,其识别:

第一数据集候选者,所述第一数据集候选者的表示属于所述第一数据流图表,并且所述第一数据集候选者匹配用于识别所述第一数据集的不同版本的一个或多个准则;以及

第二数据集候选者,所述第二数据集候选者的表示属于所述第二数据流图表,并且所述第二数据集候选者匹配用于识别所述第一数据集的不同版本的一个或多个准则,以及

用户接口,其接收指示所述第一数据集候选者和所述第二数据集候选者两者表示同一数据集的映射以及与数据流图表关联来在数据存储系统中存储所述映射,所述数据流图表向所述映射的所述第一数据集候选者和所述第二数据集候选者提供数据或从所述映射的

所述第一数据集候选者和所述第二数据集候选者接收数据。

27. 根据权利要求 26 所述的系统,其中该用户接口呈现所述第一数据集候选者和所述第二数据集候选者。

28. 根据权利要求 26 所述的系统,其中用户接口呈现数据集之间的映射的列表,所述映射根据对一个或多个准则的匹配的数量而排序。

29. 根据权利要求 28 所述的系统,其中映射的列表包括在列表中排序较高的给定数据集的版本的候选者。

30. 根据权利要求 28 所述的系统,其中准则之一被内置于映射器中。

31. 根据权利要求 28 所述的系统,其中准则之一是通过用户接口接收的。

32. 根据权利要求 28 所述的系统,其中至少一个映射指示表示所述数据集候选者之一的数据流图表的组件,并且至少一个映射指示不表示所述数据集候选者之一的数据流图表的组件。

33. 根据权利要求 32 所述的系统,其中包括多个组件的数据流图表的子图表表示所述数据集候选者之一。

34. 根据权利要求 33 所述的系统,其中该子图表包括数据组件。

35. 根据权利要求 33 所述的系统,其中该子图表包括可执行组件。

36. 根据权利要求 26 所述的系统,其中该映射器使用试探法以确定所述第一数据集候选者是否具有与所述第二数据集候选者共同的一个或多个特征。

37. 根据权利要求 36 所述的系统,其中该特征包括在所述数据集候选者之一的表示中字节和记录的数量。

38. 根据权利要求 36 所述的系统,其中该特征包括所述数据集候选者之一的表示的名称。

39. 根据权利要求 36 所述的系统,其中该特征包括所述数据集候选者之一的表示的创建的日期。

40. 根据权利要求 36 所述的系统,其中该特征包括所述数据集候选者之一的表示的数据格式。

41. 根据权利要求 26 所述的系统,其中该映射的所述第一数据集候选者和所述第二数据集候选者的至少一个属于数据管理系统已知的数据集的组。

42. 根据权利要求 26 所述的系统,其中该映射器产生所述第一数据集候选者和所述第二数据集候选者之间的格式映射。

43. 根据权利要求 26 所述的系统,其中所述映射器存储包括标识符的映射,所述标识符指向保持所述第一数据集候选者和所述第二数据集候选者的踪迹的数据管理系统中的记录。

44. 根据权利要求 26 所述的系统,其中所述映射器基于在所述第一数据集候选者或者所述第二数据集候选者中的改变更新数据集之间所存储的映射。

45. 根据权利要求 26 所述的系统,其中数据流图表中的至少一个包括代表由表示数据的流的链路相互连接的计算的节点,所述数据流图表从所述至少一个输入数据集接收数据的流并且向所述至少一个输出数据集提供数据的流。

46. 根据权利要求 26 所述的系统,其中所述第一数据流图表为数据流图表的子图表并

且所述第二数据流图表为所述数据流图表的子图表。

47. 一种用于映射存储在数据存储系统中的数据的系统,该系统包括:

数据存储系统,其存储:

第一数据流图表的规范,其从至少一个输入数据集接收数据并且向至少一个输出数据集输出数据;以及

第二数据流图表的规范,其从至少一个输入数据集接收数据并且向至少一个输出数据集输出数据,

识别部件,用于:

识别第一数据集候选者,所述第一数据集候选者的表示属于所述第一数据流图表,并且所述第一数据集候选者匹配用于识别第一数据集的不同版本的一个或多个准则;

识别第二数据集候选者,所述第二数据集候选者的表示属于所述第二数据流图表,并且所述第二数据集候选者匹配用于识别所述第一数据集的不同版本的一个或多个准则,

用于提供用户接口以接收指示所述第一数据集候选者和所述第二数据集候选者两者表示同一数据集的映射的部件;以及

用于与数据流图表关联来存储在用户接口上接收的所述映射的部件,所述数据流图表向所述映射的所述第一数据集候选者和所述第二数据集候选者提供数据或从所述映射的所述第一数据集候选者和所述第二数据集候选者接收数据。

48. 根据权利要求 47 所述的系统,其中数据流图表中的至少一个包括代表由表示数据的流的链路相互连接的计算的节点,所述数据流图表从所述至少一个输入数据集接收数据的流并且向所述至少一个输出数据集提供数据的流。

49. 根据权利要求 47 所述的系统,其中所述第一数据流图表为数据流图表的子图表并且所述第二数据流图表为所述数据流图表的子图表。

在数据管理系统中映射数据集的实例

[0001] 相关申请的交叉引用

[0002] 本发明要求 2008 年 12 月 2 日提交的美国专利申请 : 序列号 61/119164 的优先权, 其通过引用在此并入。

技术领域

[0003] 本发明涉及在数据管理系统中映射数据集的实例。

背景技术

[0004] 现代数据管理系统可以包括代表该系统的不同方面的众多组件。不那么复杂的系统常常允许数据被直接查看, 而无需用于精确视觉化目的的另外的处理。较为复杂的系统会要求另外的机制用于有目的地查看数据。由许多组件组成的复杂的数据管理系统可以按许多不同的形式存储数据并且按许多不同的方式处理数据。这些存储和处理的形式多数可以按不显然的方式而彼此相关, 而没有一种分析该关系的办法。

发明内容

[0005] 在一个大体的方面, 一种用于映射存储在数据存储系统中的数据以由计算机系统使用的方法包括 : 通过至少一个数据流图表从至少一个输入数据集接收数据的流以及通过至少一个数据流图表向至少一个输出数据集提供数据的流, 处理包括代表由表示数据的流的链路相互连接的计算的节点的数据流图表的规范 ; 识别一个或多个数据集的集合, 其中在给定集合的每个数据集匹配用于识别单个数据集的不同版本的一个或多个准则 ; 提供用户接口以接收在给定集合的至少两个数据集之间的映射 ; 以及与数据流图表关联来存储在用户接口上接收的映射, 该数据流图表向该映射的数据集提供数据或从该映射的数据集接收数据。

[0006] 在另一个大体的方面, 一种用于映射存储在数据存储系统中的数据的系统包括 : 数据存储系统, 通过至少一个数据流图表从至少一个输入数据集接收数据的流以及通过至少一个数据流图表向至少一个输出数据集提供数据的流, 存储包括代表由表示数据的流的链路相互连接的计算的节点的数据流图表的规范 ; 映射器, 其识别与数据流图表关联的一个或多个数据集的集合, 其中在给定集合中的每个数据集匹配用于识别单个数据集的不同版本的一个或多个准则 ; 用户接口, 其接收在给定集合中的至少两个数据集之间的映射 ; 以及与数据流图表关联来在数据存储系统中存储该映射, 该数据流图表向该映射的数据集提供数据或从该映射的数据集接收数据。

[0007] 在另一个大体的方面, 一种用于映射存储在数据存储系统中的数据的系统包括 : 一部件, 用于通过至少一个数据流图表从至少一个输入数据集接收数据的流以及通过至少一个数据流图表向至少一个输出数据集提供数据的流, 处理包括代表由表示数据的流的链路相互连接的计算的节点的数据流图表的规范 ; 用于识别一个或多个数据集的集合的部件, 其中在给定集合中的每个数据集匹配用于识别单个数据集的不同版本的一个或多个准

则 ;用于提供用户接口以接收在给定集合中的至少两个数据集之间的映射的部件 ;和用于与数据流图表关联来存储在用户接口上接收的映射的部件,该数据流图表向该映射的数据集提供数据或从该映射的数据集接收数据。

[0008] 在另一个大体的方面,一种存储用于映射存储在数据存储系统中的数据的计算机程序的计算机可读介质,该计算机程序包括指令,用于促使计算机来 :通过至少一个数据流图表从至少一个输入数据集接收数据的流以及通过至少一个数据流图表向至少一个输出数据集提供数据的流,处理包括代表由表示数据的流的链路相互连接的计算的节点的数据流图表的规范;识别一个或多个数据集的集合,其中在给定集合中的每个数据集匹配用于识别单个数据集的不同版本的一个或多个准则 ;提供用户接口以接收在给定集合中的至少两个数据集之间的映射 ;以及与数据流图表关联来存储在用户接口上接收的映射,该数据流图表向该映射的数据集提供数据或从该映射的数据集接收数据。

[0009] 各个方面能够包括以下特征的一个或多个。

[0010] 在用户接口上呈现该集合。

[0011] 在用户接口上呈现根据对一个或多个准则的匹配的数量而排序 (order) 的可能的映射的列表。

[0012] 可能的映射的列表包括很可能是在列表中排序较高的给定数据集的实例的候选者。

[0013] 准则之一内置于映射器中,该映射器识别一个或多个数据集的集合。

[0014] 准则之一是从用户接口接收的。

[0015] 至少一个可能的映射指示表示数据集的数据流图表的组件,并且至少一个可能的映射指示不表示数据集的数据流图表的组件。

[0016] 包括多个组件的数据流图表的子图表表示数据集。

[0017] 该子图表包括数据组件。

[0018] 该子图表包括可执行组件。

[0019] 识别数据集的一个或多个集合包括使用试探法以确定在给定集合中的数据集是否具有与另一数据集共同的一个或多个特征。

[0020] 该特征包括在数据集的表示中字节和记录的数量。

[0021] 该特征包括数据集的表示的名称。

[0022] 该特征包括数据集的表示的创建的日期。

[0023] 该特征包括数据集的表示的数据格式。

[0024] 该映射的至少一个数据集属于对数据管理系统已知的数据集的组。

[0025] 在给定集合中的数据集之间提供格式映射。

[0026] 该映射包括标识符,其指向保持数据集的踪迹的数据管理系统中的记录。

[0027] 基于在数据集中的改变更新该映射。

[0028] 本发明的各个方面能够包括以下一个或多个优点。

[0029] 通过根据版本识别准则识别数据集的集合,能够比纯粹手动操作更有效地实现在数据集的两个实例之间的匹配。另外,通过提供用户接口以接收在至少两个数据集之间的映射,该映射将比该系统是纯粹自动的更精确。

[0030] 通过以下描述以及权利要求书,本发明的其它特征和优点将变得明了。

附图说明

- [0031] 图 1 是数据流图表。
- [0032] 图 2 是数据集映射器及其相关组件的概图。
- [0033] 图 3A-3E 是由数据集映射器处理的不同情况的图。
- [0034] 图 4 是数据集映射器操作的流程图。
- [0035] 图 5 是数据集连接映射。
- [0036] 图 6 是数据集格式映射。

具体实施方式

[0037] 1 概述

[0038] 数据处理元素可以是图表的形式。基于图表的计算使用“数据流图表”实现,该“数据流图表”由指向图表示,该指向图具有在图表中表示组件(对应于存储的数据的数据存储组件或对应于可执行过程的计算组件)的顶点,和在图表中表示组件之间的数据的流的指向链路或“边缘”。数据流图表(也简单叫作“图表”)是模块化实体。每个图表可以由一个或多个其他图表组成,并且特别的图表可以是较大图表中的组件。图形开发环境(GDE)提供用户接口用于规定可执行图表和定义关于图表组件的参数。

[0039] 参考图 1,数据流图表 101 的示例包括输入组件 102,提供要由数据流图表 101 的可执行组件 104a-104j 处理的数据积聚集(collection)。例如,数据集 102 可以包括与数据库系统关联的数据记录或与交易处理系统关联的交易。每个可执行组件与由整个数据流图表 101 定义的计算的一部分关联。工作要素(如,来自数据积聚集的各个数据记录)进入组件的一个或多个输入端口,而输出工作要素(它们在一些情况中是输入工作要素,或输入工作要素的处理后的版本)一般离开该组件的一个或多个输出端口。在图表 101 中,来自组件 104e、104g 和 104j 的输出工作要素被存储在输出数据组件 102a-102c 中。

[0040] 数据集是表示特定的数据积聚集的对象(如,存储在面向对象的数据库中)。在数据流图表的系统的环境中,组件能够表示数据集。在这些情况中,图表可以按一种或多种方式与表示数据集的组件(或简称“数据集组件”)交互作用。数据集组件包括用于访问由给定数据集表示的物理数据的指令,从而图表可以使用数据集组件接受来自数据集的输入,使用数据集组件提供输出到数据集,并且在中间步骤中使用数据集组件处理数据集的数据。数据集组件可以包括与包括数据集对象的实例的给定数据集对象相关的各种类型的信息。此类系统可以具有多打、成千上百的图表和相关的数据集组件。随着此类系统的复杂性增加,不同图表和数据集组件之间的关系变得更难以管理。在系统中一个以上的数据集组件可以表示相同的数据源并且每个此类数据集组件可以与不同图表、图表子集或可执行组件关联。

[0041] 例如,在一个可能的情况中,单个数据集可以存储在与数据管理系统关联的一个以上的位置中。在该情况中,两个或多个数据源包含该相同数据的相似或相同的版本。在该系统中的两个图表可能处理该单个数据集,但是每个图表从不同数据文件、不同数据库表、或另一类型的数据集组件读取和将每个图表写入不同数据文件、不同数据库表、或另一类型的数据集组件。

[0042] 在相似的情况中,由给定数据集表示的数据(例如数据文件)可以不仅存储在一个以上的位置中,而且使用不同的数据存储格式来解释。通过以上示例,两个图表可以操作于包含相同数据仅在格式上不同的两个单独的数据文件。每个数据文件可以具有不同的数据类型,尽管包含相同数据的实例。

[0043] 在替换的情况中,一个图表可以操作于包含该数据集的实例的数据文件,而另一图表可以操作于也包含该数据集的实例的数据库表。在此种情况中,数据文件和数据库表将通常具有两个不同的数据格式。

[0044] 在另一情况中,数据管理系统可以按不同方式访问不同版本的相同数据集的每个。一个图表可以直接访问数据集的实例,诸如通过经过标准的文件输入/输出机制读取数据文件。另一图表可以通过查询外部源(诸如经由网络可用的数据仓库)检索文件。图表也可以访问经过类似的外部查询(诸如对联网数据库的查询)检索的数据库表。

[0045] 数据管理系统也可以按不同方式引用(reference)相同数据集每个的不同实例。例如,图表能够根据参数访问不同的数据位置。此类参数可以随时间指向任意数量的数据位置。如果参数在图表的执行之间改变,则操作多次的图表可以在不同场合中访问不同的位置。

[0046] 在一些情况中,在图表中数据集的表示可以不是单个组件,而是组件的积聚集,诸如图表中的“子图表”组件,其自身实现为具有多个组件的图表。该积聚集可以包括一个或多个数据集组件,并且也可以包括一个或多个可执行组件。

[0047] 所有这些情况可以潜在地对显现和分析由数据管理系统处理的数据造成问题。如果用户要求与给定数据集交互作用的各组件的统一查看,则能够使用各种办法来使可能存在的数据集的不同实例协调(reconcile)。

[0048] 一个办法是自动机制,其识别相同数据集的多个实例并且创建在它们之间的连接。可是,有些自动机制具有缺点,诸如以下的三种缺点。第一,该机制会要求数据集的每个实例按特定方式存储,诸如按照统一的命名方案和目录结构。这给该机制提供一种方法以识别和定位与数据管理系统关联的存储系统中的每个。可是,这种安排限制了数据管理系统的灵活性并且可能对该系统的某些使用过于限制。

[0049] 第二,在操作的几个情况中,该机制可能无法恰当识别相同数据集的实例以及形成正确的连接。例如,如果使用外部引用实体来访问数据集,则这很有可能,并且该自动机制未能访问该实体。类似地,如果组件根据参数列表中的独立参数访问数据集,则这很有可能,并且该机制没有办法访问或解释该参数列表。另外,如果数据集通过由一个或多个数据集组件和可执行组件组成的复杂实体(诸如子图表)表示,则这很有可能。自动机制可能无法辨别组件的什么特定组合表示特定的数据集。

[0050] 第三,该机制会形成在数据集的实例之间的冗余或不必要的连接。例如,由数据管理系统处理的一些数据集可以表示额外的数据,诸如错误内容日志。在这些数据集的实例之间的任何链接是不必要的。另外,由数据管理系统处理的一些数据集实例可能是冗余实例,诸如高速缓存数据或数据的其他临时副本。连接这种类型的数据的链接很快变为过时并且使检查该数据管理系统的用户搞糊涂。

[0051] 替换办法是这样的系统,其中用户经由用户接口手动地统一相同数据集的实例。用户不太可能错过在数据集的实例之间的本质链接,并且也不太可能产生在数据集的实例

之间的冗余或不必要的链接。可是,如果该数据管理系统具有数百或数千的组件,则用户手动产生必须的链接所需的时间量高得惊人得大。

[0052] 在部分自动的办法中,数据集映射器用来提供一些自动分析,并且按照对大且/或复杂系统的用户不高得惊人的方法使用户的一些交互作用。

[0053] 图 2 是示出在相关的主要组件之间的相互关系的示范数据集映射器 100 的一个实施例的框图。数据集映射器 100 能够分析一个或多个图表 180、180a、180b、180c 的集合。每个图表与一个或多个数据集组件 182、182a、182b 关联,其中每个数据集组件可以对应于数据文件、数据库表、子图表或表示数据集的另一类型的组件。映射器 100 分析各图表,目的是在包含相同数据集 186 的实例的数据集组件之间形成链接。映射器 100 根据内置规则 110、用户定义规则 120 和试探法 130 的组合处理每个数据集组件以确定数据集组件 182 是否可以包含表示数据管理系统 170 已知的数据源 176、176a、176b 的几个数据集之一的实例。映射器 100 将该信息传递给用户接口 160,其允许用户 162 选择对应于数据集组件 182 的适当的数据集(如果有的话)。例如,用户接口 160 呈现基于匹配于一个或多个准则(用于识别单个数据集的不同的版本或实例)的可能的候选映射的列表。下面将更详细地描述此类准则的示例,其包括基于内置规则、用户定义规则和试探法的准则。该列表可以根据对一个或多个准则的匹配的量化来排序(如,更有可能成为给定数据集的实例的候选者在列表中排序较高)。映射器 100 然后产生数据集链接映射 140,其指示数据集组件 182 包含表示数据源 176 的数据集的实例。

[0054] 另外,数据集组件 182 可以具有不同于对应链接的数据源 176 的格式 174 的数据格式 184。取决于数据管理系统 170 的要求,用户可以选择为数据集的全部实例建立单个数据格式。该系统存储关于每个数据源 176、176a、176b 的格式 174、174a、174b。替换地,用户可以选择产生在数据集组件 182 的格式 184 和对应数据源 176 的建立格式 174 之间的可选的映射 142。可选的数据格式映射 142 允许系统 170 保持关于数据集的每个实例的数据类型的信息。

[0055] 映射器 100 也使得用户能够指示在可执行组件和单个数据集组件(它可以不具有其他的连接)之间的链接。例如,数据集组件可以对应于仅具有一个读取器的源数据集或仅具有一个写入器的目标数据集。如果数据集对象已经存在于系统中并且具有其他相关元数据,诸如正确记录格式、文件、数据简档等,该链接使能数据集组件映射到正确的数据集。

[0056] 2 映射过程

[0057] 映射器 100 能够处理在复杂的数据管理系统中出现的共同情况。在第一情况中,如图 3A 所示,一个图表 210 提供数据集组件 212 作为输出,而另一图表 220 接受一不同的数据集组件 222 作为输入。每个数据集组件包含相同数据集 216 的实例。该数据集可以和表示数据管理系统已知的数据源 176 的数据集相同。另外,第一数据集组件 212 具有可以和属于第二数据集组件 222 的格式相同的数据格式 214,或,替换地,第二组件可以具有不同的格式 224。映射器 100 能够识别第二数据集组件 222 作为由第一数据集组件 212 表示的数据集 216 的实例并且产生适合的链接映射 140。

[0058] 在第二情况中,如图 3B 所示,图表 230 使用对外部源 239 的外部引用 238 与外部数据集组件 232 关联。外部数据集组件 232 具有数据格式 234 并且是数据集 236 的实例。如第一情况一样,由外部数据集组件表示的数据集 236 可以是表示数据管理系统 170 已知

的数据源 176 的数据集。映射器 100 能够识别该外部数据集组件 232 作为另一数据集的实例并且产生适合的链接映射 140。

[0059] 在第三情况中,如图 3C 所示,图表 240 使用参数列表 247 的参数 238 与数据集组件 242 关联。引用的数据集组件 242 具有数据格式 244 并且是数据集 246 的实例。如第一和第二情况一样,由引用的数据集组件表示的数据集 246 可以是表示数据管理系统 170 已知的数据源 176 的数据集。映射器 100 能够识别该引用的数据集组件 242 作为另一数据集的实例并且产生适合的链接映射 140。

[0060] 在第四情况中,如图 3D 所示,图表 250 使用对外部源 259 的外部引用 258 与外部组件 251 关联。外部组件 251 不是数据集组件,而是另一类型的组件,诸如可执行组件。映射器 100 能够识别该可执行组件 251 为不适用于数据集链接映射过程。

[0061] 在第五情况中,如图 3E 所示,图表 260 与自身由几个组件组成的子图表组件 263 关联。这些组件包括至少一个数据集组件 262,以及在该示例中,包括一个或多个可执行组件 261a、261b、261c。在这种情况下,作为单个实体的子图表 263 表示至少一个数据集。其他示范的子图表可以包括多个数据集组件,和任何数量(包括零个)的可执行组件。另外,该子图表 263 具有多个输出 265a、265b。每个输出能够提供数据集的不同实例到接收该输出的组件。另一示范的子图表还可以具有任何数量的输入。再一示范的子图表可以不具有对应于各个数据集的输入或输出。对于子图表确实表示至少一个数据集的情况,映射器 100 能够识别该子图表 263 作为至少一个数据集的实例并且产生至少一个适合的链接映射 140。

[0062] 在图 4 中示出映射器的操作序列的示例。在步骤 302 中,映射器首先识别在与图表关联的要素中哪个要素表示数据集。一般,图表将具有一个或多个输入和输出,并且每个输入和每个输出可以是数据集的实例。每个图表也可以在一些中间步骤中处理数据集的实例。结果,每个图表可以连接到能够作为数据集候选者的多个组件。在有些情况中,数据管理系统具有关于某些组件的特征的信息,包括有关该组件是否表示数据集的信息。在那些情况下,在步骤 304 中映射器将潜在的数据集组件添加到数据集候选者的表中。在有些情况下,组件可以是由包括数据集组件和可执行组件的多个组件组成的子图表。子图表可以表示数据集的至少一个实例。因此,映射器编译全部此类子图表的列表并且作为步骤 304 的一部分将它们添加到数据集候选者的表中。在其他情况中,组件的性质可能不可用于数据管理系统。该组件可以通过对外部实体的引用而被访问,其中该引用可以是对数据库表的查询、指向因特网服务器的统一资源定位符、参数列表中的参数、或另外类型的引用。在这些情况下,映射器一般不具有这样的手段,通过它映射器可以独立访问由引用指向的实体。因此,映射器编译全部此类引用的列表并且作为步骤 304 的一部分将它们添加到数据集候选者的表中。

[0063] 接下来,在步骤 306 中,对于给定的数据集候选者,映射器产生数据集候选者能够映射到的已知数据集的列表。该映射器使用用户定义规则、内置规则和试探法的组合来评估哪个已知数据集能够映射到数据集候选者。

[0064] 接下来,在步骤 308,用户然后选择对应于数据集候选者的已知数据集。如果建议的已知数据集全不是正确的匹配,则用户也可以访问全部已知数据集的整个列表。此外,用户能够指示数据集候选者不是数据集。例如,对远程服务器的引用可以是对远程可执行过

程（它不是数据实体）的调用。作为另一示例，数据集候选者可以表示数据，但是它可以是属于数据管理系统的类型的数据，诸如错误日志。在这种情况下，用户可以向用户接口指示该数据将在映射过程中被忽略。

[0065] 接下来，在步骤 310 中，用户识别新映射的数据集的数据格式。该系统可以具有数据格式模板的集合，能够选择其中一个模板。替换地，用户能够在用户接口中产生新的数据格式。

[0066] 接下来，在步骤 312 中映射器使用该信息产生关于数据集候选者的链接映射，以及可选的，格式映射。

[0067] 接下来，映射器将下一个数据集候选者提供给用户用于在步骤 308、310 和 312 的另一迭代中的链接产生，除非映射器已经处理全部的数据集候选者。

[0068] 接下来，在步骤 314 中，用户查看与数据管理系统关联的组件以基于在组件之间的新链接确保在图表和数据集组件之间的关联的显现是精确的。在步骤 316 中，用户具有对链接和格式映射做出任何调整的选项。

[0069] 最后，在步骤 318 中，映射器向数据管理系统传递该链接和格式映射。该映射可以与一个或多个图表一起存储，或存储在与数据管理系统关联的单独的存储实体中，或通过另外的手段存储。

[0070] 3 数据集映射维持

[0071] 映射器 100 能够处理可能出现的多个情况，其影响数据集链接的完整性。

[0072] 第一情况包括当新组件被添加到数据管理系统 170 时识别新的数据集候选者。在这种情况下，映射器 100 分析每个组件并且向用户呈现可能的链接。映射器 100 能够操作于任何新的组件以产生需要的合适链接。

[0073] 第二情况包括当数据管理系统 170 随时间改变时维持现有的链接。例如，数据集的新实例可以在与系统关联的图表的正常操作的期间开始存在。作为另一示例，数据集可能已经改变其身份，诸如其名称或系统中的位置。作为另一示例，数据集可以被整个删除。作为再一示例，数据集候选者可能在前一轮链接产生中被遗漏，从而链接的积聚集不完整。映射系统的用户接口 160 允许用户 162 修改现有的链接以弥补不完整或过时的任何映射。

[0074] 第三情况包括自动更新关于总是遵循已知模式的数据集引用的链接。例如，图表可以处理在参数列表 247 中引用的数据集。此类参数列表可以随时间改变。如果参数列表遵循数据管理系统已知的标准格式，则映射器可以识别在参数列表中的改变并且相应地更新现有的链接。

[0075] 4 数据集链接映射

[0076] 如图 5 所示，数据集链接映射 140 包含组件名称 402、数据集名称 404、数据集类型 406、格式 408、主要数据集位置 410 和标记 412。组件名称 402 是表示数据集的实例的数据集组件或子图表。数据集名称 404 是指向由该组件表示的数据集的标识符。数据集类型 406 指示在例如数据文件、或数据库表、或另外类型的情况下该数据集的实例落入的类别。格式 408 是数据集的实例用来表示其数据的格式或安排。主要数据集位置 410 是标识符，其指向保持该数据集的踪迹的数据管理系统中的记录。最后，标记 412 指示该数据集的实例是否应该被忽略，例如，如果用户已识别数据集的实例为不适用于该数据管理系统并且应该被从链接的集合中剔除。

[0077] 5 内置规则

[0078] 映射器 100 具有内置规则 110 的集合,其根据数据管理系统的标准惯例而操作。如果数据集组件遵循内置规则 110,则映射器能够以最高的精确度识别对应于数据集组件的数据集。在规则的一个示范实施中,包含数据集候选者的外部引用的数据库表必须按照由数据管理系统使用的标准化目录结构而安置在永久存储器中。另外,根据参数访问外部引用的数据集组件的图表必须使用数据管理系统也能够访问和解析的参数。另外,数据集组件的格式必须在永久存储器中可用并且可由数据管理系统访问。取决于数据管理系统,其他内置规则也是可能的。

[0079] 6 用户定义的规则

[0080] 除了映射器用来识别数据集候选者的内置规则外,映射器 100 也具有可选的用户定义规则 120 的积聚集。这些规则 120 可以由用户使能或禁止,这取决于哪个适用于用户的特定数据管理系统。在一个示范实施中,映射器具有六个用户定义的可选规则。如果在名称中一些信息模糊该表的标识,诸如关于定义该表的用户的信息,则映射器可以忽略在数据库表的名称中的一些信息。另外,映射器能够从数据库表的名称中消除该信息。另外,映射器可以忽略数据文件的特定目录,它们已知为包含不属于与数据管理系统关联的数据集的数据。此类目录可以是数据文件类型或数据文件扩展。另外,映射器能够解析对参数列表的特定参数的引用并且以参数自身的名称取代该引用。另外,映射器能够整个消除对参数的引用。用户也能够产生映射器要遵循的其他规则。

[0081] 7 试探法

[0082] 除了遵循内置和用户定义的规则来评估数据集候选者外,映射器 100 也使用试探法 130 的集合。试探法 130 允许映射器分析给定数据集组件的特征并且将那些特征与已知的数据集比较。与已知的数据集具有类似特征的数据集组件很可能是数据集的实例。在一个示范实施中,映射器使用两个试探法。一个试探法是给定数据集组件的数据的特征。例如,如果与数据集组件关联的数据具有和与已知的数据集关联的数据相同数量的字节和记录,则该数据集组件很可能是那个数据集的实例。另外,如果数据集组件具有类似于已知的数据集的名称或产生的时期,则该数据集组件很可能是那个数据集的实例。第二试探法是数据集组件的数据格式。如果数据集组件与已知的数据集共享数据格式,则该数据集组件很可能是数据集的实例。该试探法在其中多个不同的数据集使用相同数据格式的情况下不那么可靠。

[0083] 8 数据集格式和映射

[0084] 表示数据源的每个数据集具有关联的数据格式,其对数据集的每个要素指示该要素代表什么类型的数据。例如,数据库表的数据格式指示在给定记录中每个字段的数据类型。数据管理系统 170 对于表示数据源 176、176a、176b 的每个数据集保持单个数据格式 174、174a、174b。

[0085] 如果映射器 100 已遇到表示新的数据集 186 的数据集组件 182,则映射器 100 基于数据集组件 182 的数据格式 184 产生将由数据管理系统存储的对应的数据格式。

[0086] 在其中数据集组件 183 表示代表数据源 176 的已知的数据集的一些情况中,数据集组件 182 具有不同于代表数据源 176 的已知的数据集的数据格式 174 的数据格式 184。数据管理系统 170 将代表数据源 176 的数据集作为单个实体处理,独立于数据集可能存在

的实体的数量。结果,当这些情况出现时数据管理系统 170 依靠映射器 100 来统一不同的格式 174、184。在一个实施中,映射器能够依靠于用户和数据管理系统的要求以四种不同方法的一种来应对每个情况。用户 162 能够选择统一的四种方法的任何一种用于每个情况。

[0087] 在统一的第一方法中,映射器 100 使用数据集组件 182 的数据格式 184 作为数据集的主要数据格式并且相应地更新数据管理系统 170。

[0088] 在统一的第二方法中,映射器 100 使用现有数据集的数据格式 174 作为数据集的主要数据格式并且相应地更新数据管理系统 170。

[0089] 在统一的第三方法中,映射器 100 保持两种数据格式,并且产生在每个数据格式的字段之间的映射 142。如图 6 所示,数据格式映射 142 指示数据集格式 510 的哪个字段 512a、512b、512c 对应于数据集实例(如,数据集组件)的格式的哪个字段 522a、522b、522c。

[0090] 在统一的第四方法中,映射器产生能够充当任一数据格式的新的统一的数据格式。

[0091] 9 一般计算机实施

[0092] 以上描述的数据集映射方法能够使用用于在计算机上执行的软件来实现。例如,软件形成在一个或多个编程或可编程的计算机系统(其可以具有各种架构,诸如分布式的、客户端/服务器或网络的)中执行的一个或多个计算机程序中的过程,每个计算机系统包括至少一个处理器、至少一个数据存储系统(包括易失性和非易失性存储器和/或存储元件)、至少一个输入设备或端口、以及至少一个输出设备或端口。该软件可以形成较大程序的一个或多个模块,例如,其提供与数据流图表的设计和配置相关的其他服务。图表的节点和要素能够实现为存储在计算机可读介质中的数据结构或实现为遵循存储在数据仓库中的数据模型的其他组织的数据。

[0093] 软件可以提供在存储介质中,诸如 CD-ROM,可通过通用或专用可编程计算机读取或经过网络的通信介质传递(按可传播信号编码)到执行处的计算机。全部的功能可以在专用计算机上执行,或使用诸如协处理器的专用硬件执行。软件可以按分布方式实现,其中由软件规定的计算的不同部分由不同计算机执行。每个此类计算机程序优选地存储于或下载到存储媒介或设备(如,固态存储器或媒介,或磁或光媒介)中,可由通用或专用可编程计算机读取,以用于当通过计算机系统读取存储媒介或设备以执行这里所述的过程时配置和操作该计算机。本发明的系统也可以考虑实现为计算机可读存储介质,配置为计算机程序,其中存储介质的配置使得计算机系统按特定和预定义的方式操作以执行这里所述的功能。

[0094] 已经描述了本发明的多个实施例。然而,将理解各种修改可以被做出而不脱离本发明的精神和范围。例如,上述的一些步骤可以是顺序独立的,因此能够按照不同于所述的顺序来执行。

[0095] 应该理解前述说明的意图是说明性的而非限制本发明的范围,该范围由所附权利要求的范围定义。例如,上述的多个功能步骤可以按照不同的顺序执行而不会本质上影响整个处理。其他实施例处于以下权利要求的范围中。

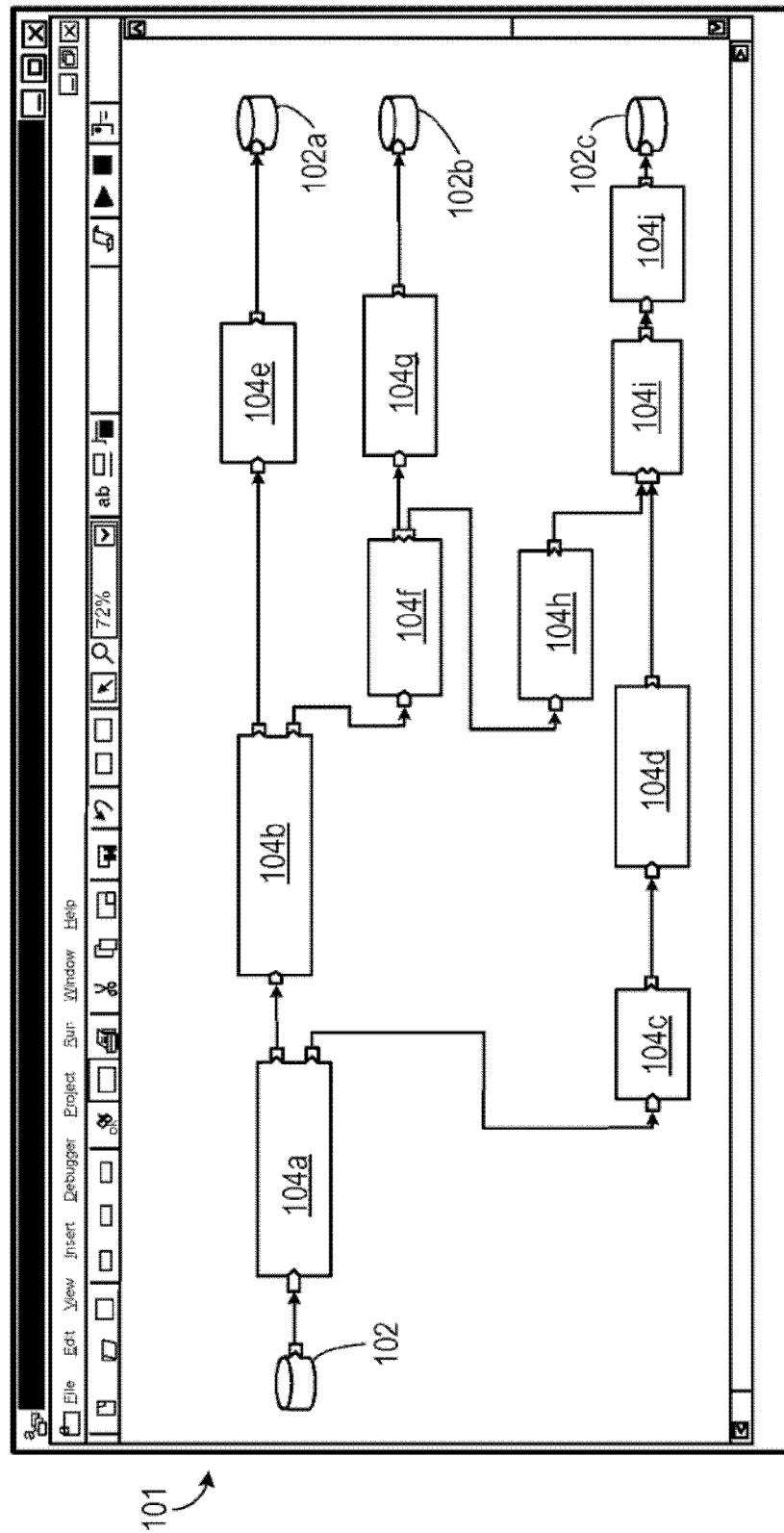


图 1

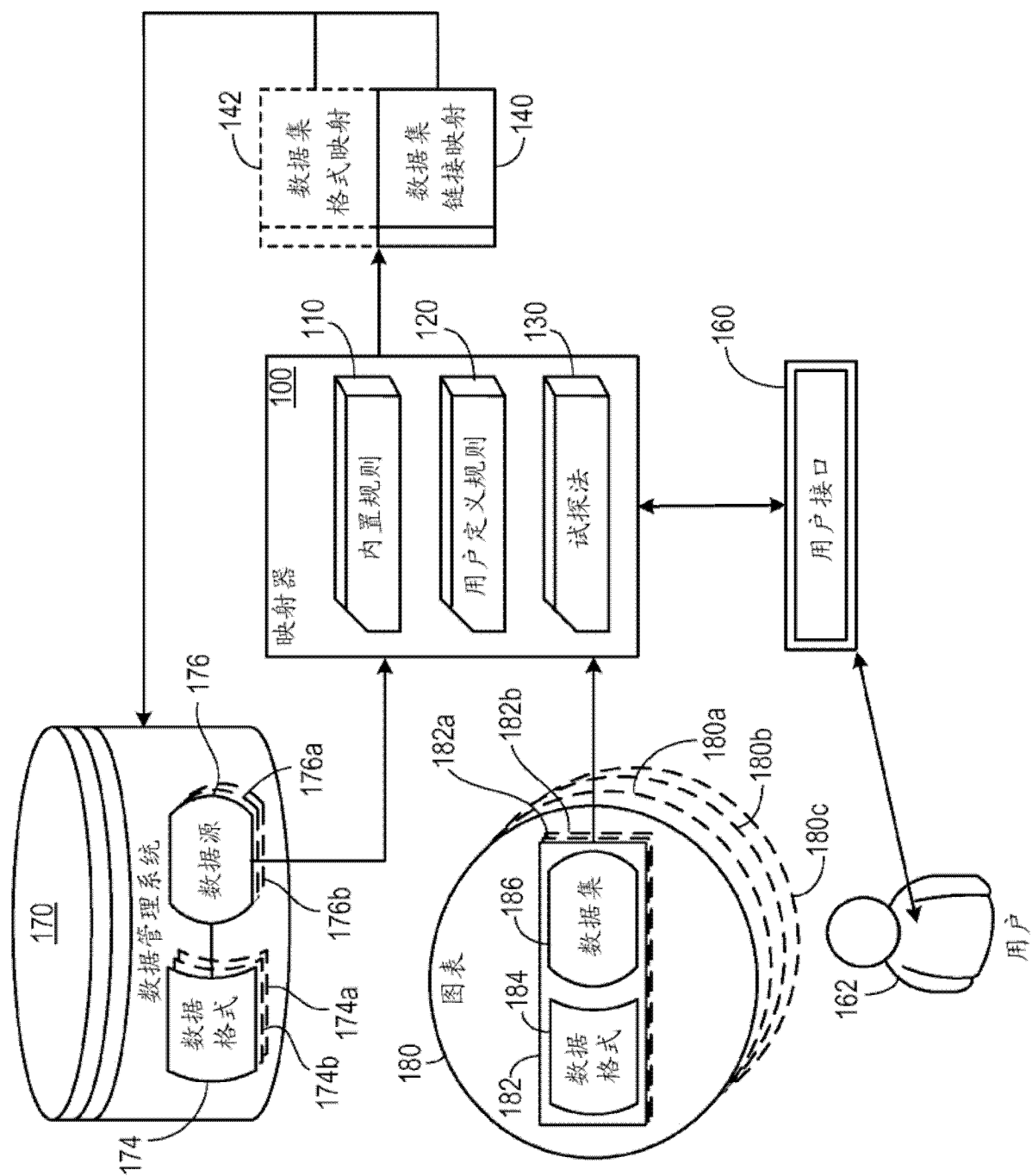


图 2

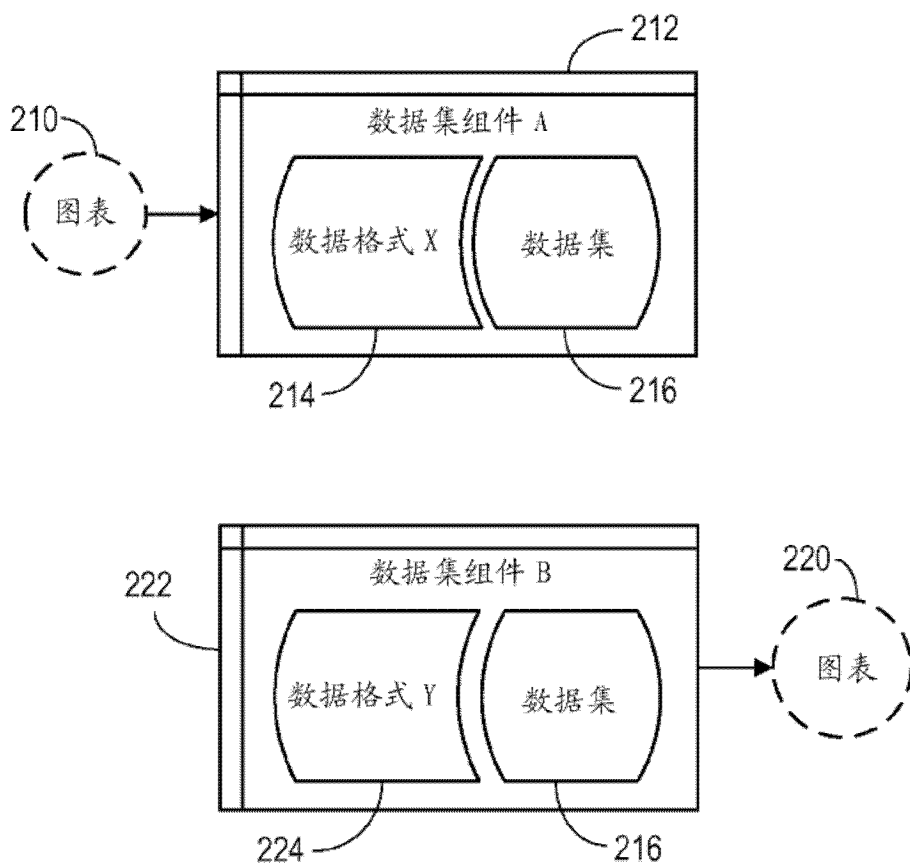


图 3A

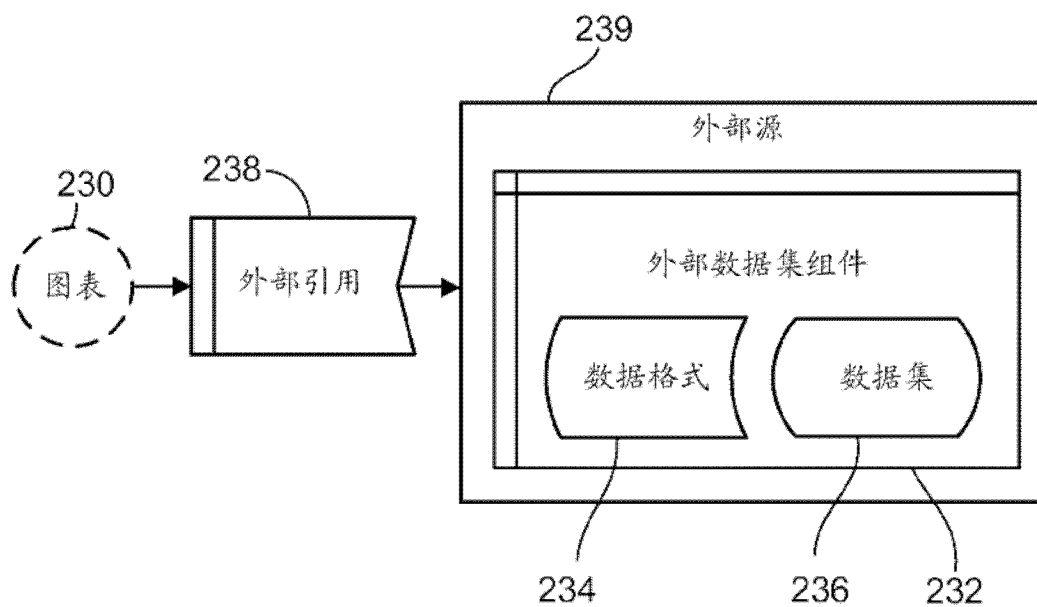


图 3B

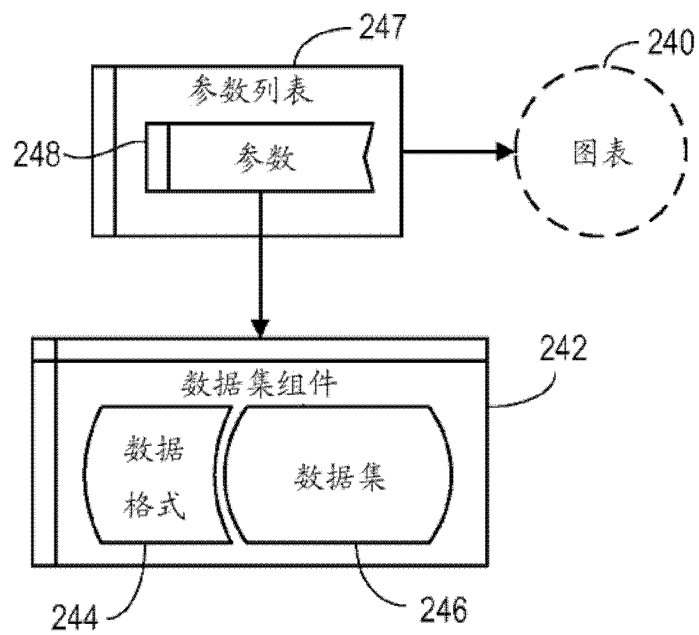


图 3C

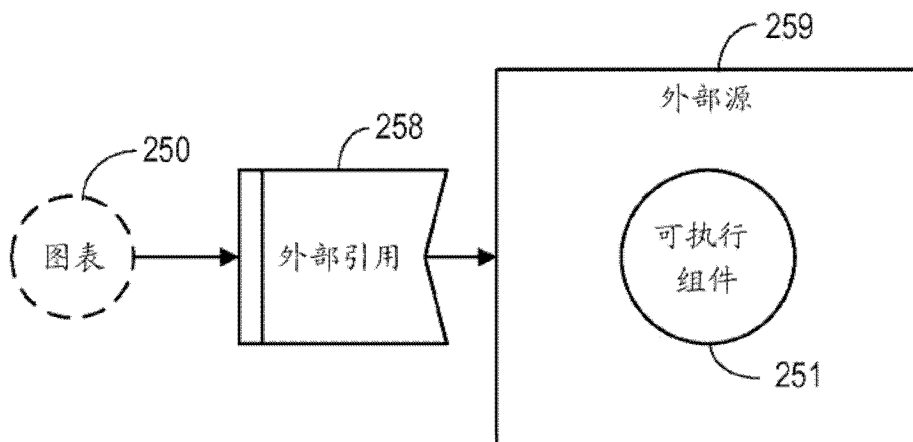


图 3D

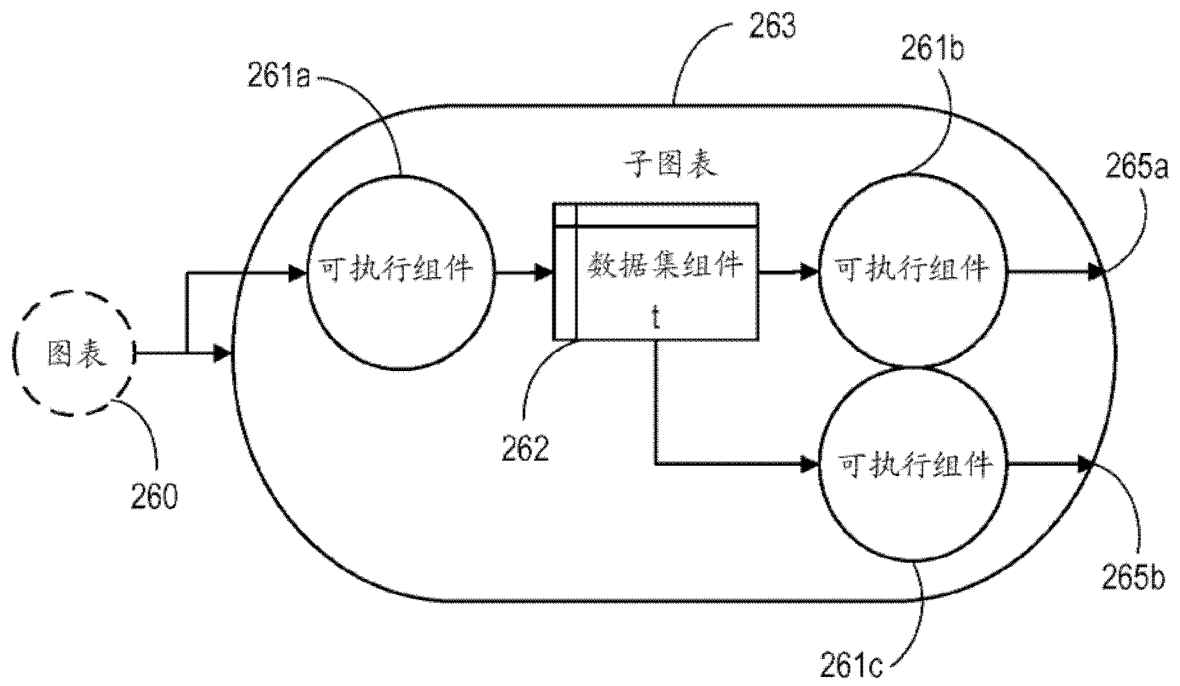


图 3E

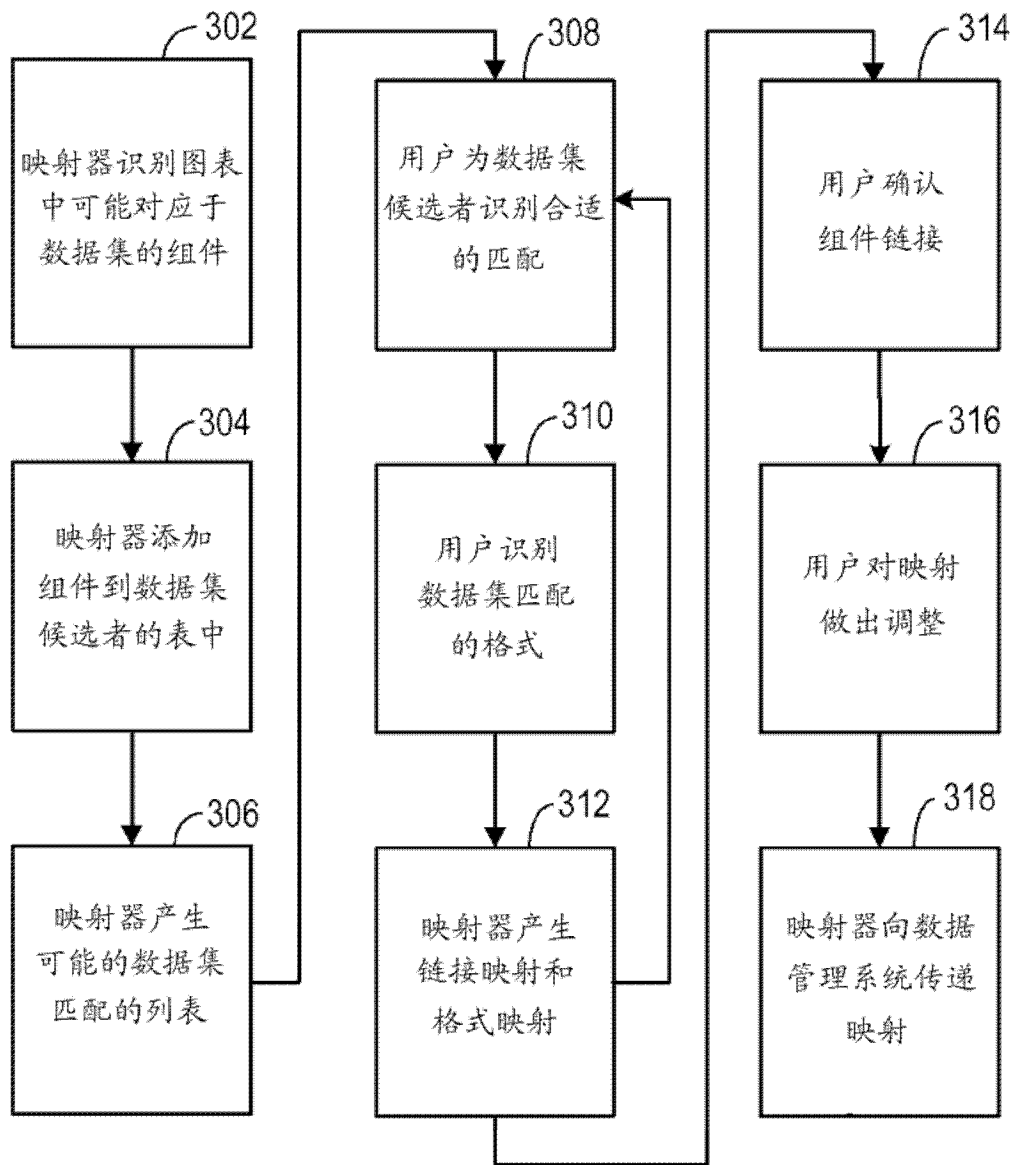


图 4

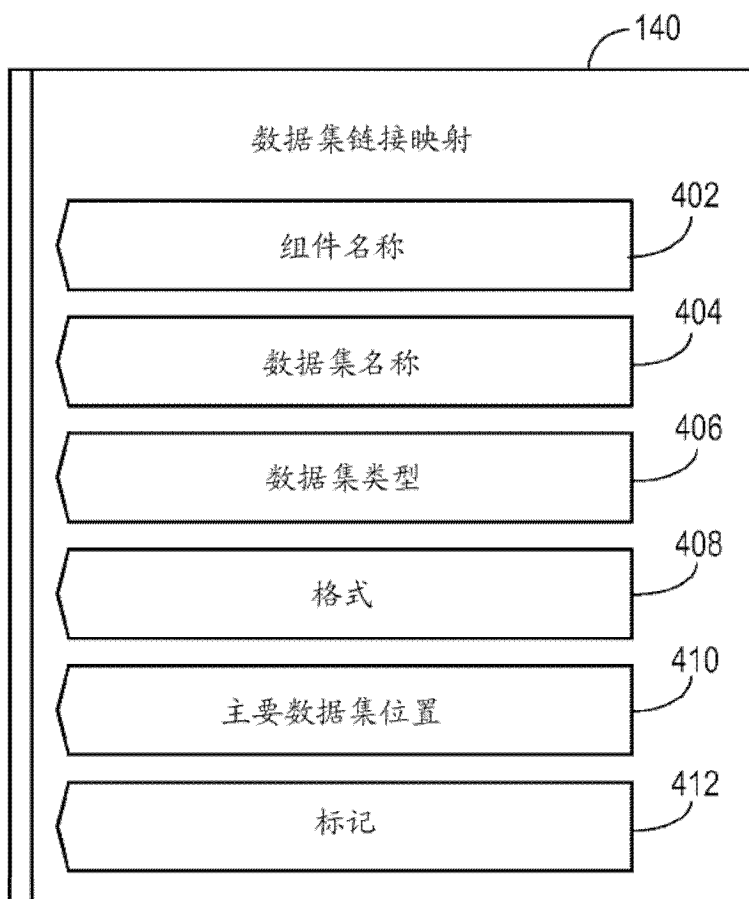


图 5

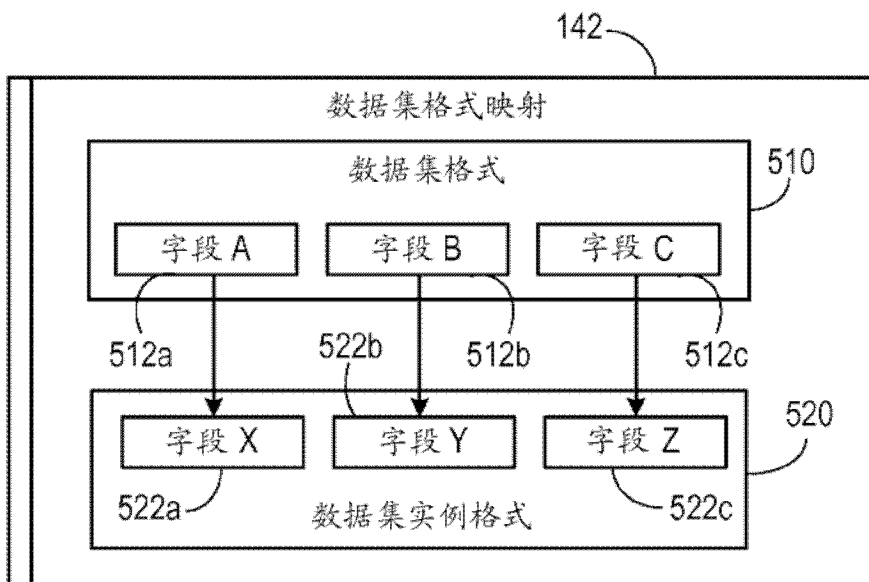


图 6