(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 103403677 B
(45)授权公告日 2017.08.11

(21)申请号 201280011660.9
(22)申请日 2012.01.05
(65)同一申请的已公布的文献号
申请公布号 CN 103403677 A
(43)申请公布日 2013.11.20
(30)优先权数据
61/430,625 2011.01.07 US
(85)PCT国际申请进入国家阶段日
2013.09.04
(86)PCT国际申请的申请数据
PCT/US2012/020334 2012.01.05
(87)PCT国际申请的公布数据
WO2012/094496 EN 2012.07.12
(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 A.F.罗伯茨
(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 9/44(2006.01)
G06F 11/36(2006.01)
(56)对比文件
US 7406424 B2,2008.07.29,
US 2005/0175341 A1,2005.08.11,
US 2006/0085532 A1,2006.04.20,
CN 102138139 A,2011.07.27,
CN 102232212 A,2011.11.02,
审查员 辛小霞

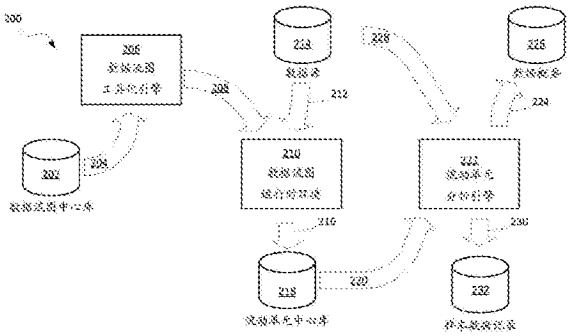
权利要求书2页 说明书11页 附图11页

(54)发明名称

流动分析工具化

(57)摘要

本发明提供了用于流动分析的方法、系统、和装置,包括编码在计算机存储介质上的计算机程序。在一个方面中,该方法包括修改(206)数据流图,该数据流图包括连接至少一个入口点和至少一个出口点的多条路径,该修改数据流图包括将把流动单元加入数据记录中和从数据记录中除去流动单元的部件加入数据流图中,每个流动单元标识数据记录穿过的一段路径。该方法还包括根据使用所修改数据流图处理多个数据记录获得的流动单元识别(222)执行路径。该方法还包括确定(230)多个数据记录的子集(232),其中该子集代表所选一组执行路径。



CN 103403677 B

1.一种计算机实现的方法,其包括:

修改数据流图,该数据流图包括连接至少一个入口点和至少一个出口点的多条路径,该修改数据流图包括:

将把流动单元加入数据记录中和从数据记录中除去流动单元的部件加入数据流图中,其中,每个流动单元利用标识以下的信息来标记指定数据记录:(i)经所述数据流图的路径中的被所述指定数据记录穿过的一段、以及(ii)在所述指定数据记录依赖于一个或多个其它数据记录的情况下所述指定数据记录所依赖的一个或多个其它数据记录;

针对利用修改后的数据流图所处理的数据记录,

基于标记所述数据记录的一个或多个流动单元来生成记录谱系,所述谱系指定(i)所述数据流图的所述多条路径中的哪条路径被所述数据记录穿过、以及(ii)在所处理的数据记录依赖于一个或多个其它数据记录的情况下所处理的数据记录所依赖的一个或多个其它数据记录;

基于所生成的记录谱系,标识所述数据记录经所述修改后的数据流图的执行路径,所述修改后的数据流图包括连接所述至少一个入口点和所述至少一个出口点的所述多条路径;以及

基于经所述修改后的数据流图的所述执行路径中的所选择的一组执行路径,确定所述多个数据记录的子集,该子集的数据记录已穿过所述所选择的一组执行路径,其中所述修改后的数据流图包括连接所述至少一个入口点和所述至少一个出口点的所述多条路径。

2.如权利要求1所述的方法,其中识别执行路径包括确定加入数据记录中的一组流动单元。

3.如权利要求1所述的方法,进一步包括识别所述多条路径中不在所述执行路径中的未用路径。

4.如权利要求1所述的方法,进一步包括使用所述数据流图处理多个数据记录的所述子集。

5.如权利要求1所述的方法,其中处理多个数据记录包括将第一流动单元加入所述多个数据记录中的数据记录中。

6.如权利要求5所述的方法,其中处理多个数据记录包括将第二流动单元加入该数据记录中,以及将所述第一流动单元加入所述第二流动单元中。

7.如权利要求1所述的方法,其中识别执行路径包括:

从使用修改后的数据流图处理的多个数据记录中除去流动单元;以及

分析所除去的流动单元以便为每个数据记录确定执行路径。

8.如权利要求7所述的方法,其中确定多个数据记录的所述子集包括识别具有所述所选择的一组执行路径中的一条执行路径的数据记录。

9.一种计算机实现的系统,其包括:

编程成执行包括如下的操作的一台或多台计算机:

修改数据流图,该数据流图包括连接至少一个入口点和至少一个出口点的多条路径,该修改数据流图包括:

将把流动单元加入数据记录中和从数据记录中除去流动单元的部件加入数据流图中,其中,每个流动单元利用标识以下的信息来标记指定数据记录:(i)经所述数据流图的路径

中的被所述指定数据记录穿过的一段、以及(ii)在所述指定数据记录依赖于一个或多个其它数据记录的情况下所述指定数据记录所依赖的一个或多个其它数据记录；

针对利用修改后的数据流图所处理的数据记录，

基于标记所述数据记录的一个或多个流动单元来生成记录谱系，所述谱系指定(i)所述数据流图的所述多条路径中的哪条路径被所述数据记录穿过、以及(ii)在所处理的数据记录依赖于一个或多个其它数据记录的情况下所处理的数据记录所依赖的一个或多个其它数据记录；

基于所生成的记录谱系，标识所述数据记录经所述修改后的数据流图的执行路径，所述修改后的数据流图包括连接所述至少一个入口点和所述至少一个出口点的所述多条路径；以及

基于经所述修改后的数据流图的所述执行路径中的所选择的一组执行路径，确定所述多个数据记录的子集，该子集的数据记录已穿过所述所选择的一组执行路径，其中所述修改后的数据流图包括连接所述至少一个入口点和所述至少一个出口点的所述多条路径。

10.如权利要求9所述的系统，其中识别执行路径包括确定加入数据记录中的一组流动单元。

11.如权利要求9所述的系统，进一步包括识别所述多条路径中不在所述执行路径中的未用路径。

12.如权利要求9所述的系统，进一步包括使用所述数据流图处理多个数据记录的所述子集。

13.如权利要求9所述的系统，其中处理多个数据记录包括将第一流动单元加入所述多个数据记录中的数据记录中。

14.如权利要求13所述的系统，其中处理多个数据记录包括将第二流动单元加入该数据记录中，以及将所述第一流动单元加入所述第二流动单元中。

15.如权利要求9所述的系统，其中识别执行路径包括：

从使用修改后的数据流图处理的多个数据记录中除去流动单元；以及

分析所除去的流动单元以便为每个数据记录确定执行路径。

16.如权利要求15所述的系统，其中确定多个数据记录的所述子集包括识别具有所述所选择的一组执行路径中的一条执行路径的数据记录。

流动分析工具化

[0001] 交叉引用相关申请

[0002] 本申请要求2011年1月7日提交、发明名称为“Flow Analysis Instrumenta-tion(流动分析工具化)”的美国临时申请第61/430,625号的优先权,特此通过引用并入其全部内容。

技术领域

[0003] 本发明涉及流动分析。

背景技术

[0004] 数据流图用于对数据进行操作。将数据供给数据流图。数据流图对数据进行一系列操作。在一些情形下,对数据进行的一系列操作可以随数据记录而变。

[0005] 将小组的数据记录用于测试数据流图;但是,选择一组数据记录可能是困难的,因为所选的该组数据记录可能代表不了生产环境下的数据记录。

发明内容

[0006] 本说明书描述与流动分析有关的技术。

[0007] 一般说来,描述在本说明书中的主题的一个方面可以用包括修改数据流图的动作的方法具体化,该数据流图包括连接至少一个入口点和至少一个出口点的多条路径。该修改数据流图包括将把流动单元加入数据记录中和从数据记录中除去流动单元的部件加入数据流图中,每个流动单元标识数据记录穿过的一段路径。该方法还包括根据使用所修改数据流图处理多个数据记录获得的流动单元识别执行路径的动作。该方法还包括确定多个数据记录的子集的动作,其中该子集代表所选一组执行路径。这些和其他实施例每一个都可选地可以包括一种或多种如下特征。识别执行路径可以包括确定加入数据记录中的一组流动单元。该特征还可以包括识别多条路径中不在执行路径中的未用路径。该特征还可以包括使用数据流图处理多个数据记录的子集。处 理多个数据记录可以包括将第一流动单元加入多个数据记录中的数据记录中。处理多个数据记录可以包括将第二流动单元加入该数据记录中,以及将第一流动单元加入第二流动单元中。识别执行路径可以包括从使用所修改数据流图处理的多个数据记录中除去流动单元,以及分析所除流动单元以便为每个数据记录确定执行路径。确定多个数据记录的子集可以包括识别具有一条所选执行路径的数据记录。

[0008] 可以实现描述在本说明书中的主题的特定实施例,以便获得一种或多种如下好处。可以简化数据流图的调试。可以选择使数据流图得到充分锻炼的数据记录的样本集合。可以随着各个记录流过图形跟踪它们。在附图和下面的描述中展示了描述在本说明书中的主题的一个或多个实施例的细节。该主题的其他特征、方面、和好处可以从该描述、附图、和权利要求书中明显看出。

附图说明

[0009] 图1例示了通过数据流图的执行路径；

[0010] 图2例示了可以确定数据记录的记录谱系(lineage)的示范性环境；

[0011] 图3例示了使用流动单元跟踪通过所修改数据流图的记录的例子；

[0012] 图4例示了消耗流动单元的例子；

[0013] 图5例示了使用流动单元跟踪通过数据流图的路径的例子；

[0014] 图6例示了识别导致输出数据记录的产生的输入数据记录的例子；

[0015] 图7例示了修改数据流图的数据源部件以便将流动单元加入数据记录中的例子；

[0016] 图8例示了修改有多个输出端口的部件以便将流动单元加入数据记录中的例子；

[0017] 图9例示了修改数据宿以便处理流动单元的例子；

[0018] 图10例示了跨过多个数据流图地使用流动单元的例子；以及

[0019] 图11例示了流动分析的示范性过程。

[0020] 各种附图中的相同标号和名称指示相同元件。

具体实施方式

[0021] 一般说来，流动分析允许更全面地理解一组数据记录内值的分布、数据 记录之间的关系、和处理数据记录以产生输出记录的方式。

[0022] 图1例示了通过数据流图的执行路径。数据流图102包括可以从入口点104(例如，数据源)到出口点106，112(例如，数据宿)处理数据记录的多条路径，例如，路径108和路径110。在该例子中，路径108从入口点104通到出口点106。路径110从相同入口点104开始但分支到出口点112。

[0023] 一般说来，数据流图由一些部件和识别数据记录在这些部件之间的流动的链路组成。这些部件包括数据源、数据宿、和用于处理的部件。数据源可以提供进入数据流图的入口点，以及可以读取通过该图形处理的一组数据记录。例如，数据源可以包括关系数据库中的表格或文件系统上的文件。数据源从表格或文件中读取记录并创建数据记录。数据宿可以提供从数据流图中出来的出口点，以及一旦数据流图完成了处理，就可以存储输出记录。数据源和数据宿可以包括，例如，关系数据库的表格或存储在文件系统上的文件。数据流图可以在计算机112或其他类型的计算机设备上执行。在其他实现中，数据流图的执行可以分配给多个计算设备。

[0024] 在一些实现中，这些部件可以包括输入端口和输出端口。这些链路将第一部件的输出端口与第二部件的输入端口连接。一些部件可以具有多个输入和输出端口。数据记录可以从入口点航行到出口点的一系列部件和链路被称为路径(例如，路径108，110)。

[0025] 流动分析是跟踪数据记录流过一个或多个数据流图的过程。流动分析使得可以在调试，测试和剖析(profiling)的领域中调试，测试和剖析一组新的应用程序。对于调试，流动分析使得可以在通过图形处理各个数据记录时跟踪各个数据记录。用户可以标记一个或多个记录，或停止在断点上，以及图形开发环境跟踪指定记录通过图形的路径，包括识别依赖于指定记录的任何记录和指定记录所依赖的任何记录。开发者可以识别可能呈现难以预料结果的输出数据记录，观看用于创建数据记录的输入数据记录，以及跟踪那些输入数据

记录以确定数据流图可能表现得出人意外的地方。

[0026] 对于测试，流动分析使用户能够生成只包含通过特定路径的记录的输入数据的子集。通过根据通过数据流图的特定路径选择输入数据，可以保护所处理数据记录的引用完整性。

[0027] 对于剖析，流动分析使用户能够创建将记录分类成群的图形，然后从这些类别中的记录所依赖的输入数据集中生成记录的子集。例如，一个图形可 以通过居住地和产品类别将输入的顾客和交易分成群，然后分解计算的输出记录据此落在“有利可图”的输出数据宿中还是落在“无利可图”的输出数据宿中的顾客和交易记录。

[0028] 图2例示了可以确定数据记录的谱系的示范性环境。运行在计算机，例如，来自图1的计算机112上的示范性系统200包括数据流图中心库202。如过程箭头204所表示，数据流图工具化引擎206从数据流图中心库202中获取数据流图。数据流图工具化引擎206修改数据流图，以便使数据流图能够随着如下面所讨论，通过修改的图形处理数据记录，跟踪它们的记录级谱系。数据流图工具化引擎206可以是，例如，运行在计算机上的进程。

[0029] 在一些实现中，数据流图工具化引擎206将使数据记录能够流过数据流图以便加以跟踪的处理部件加入数据流图中。例如，附加处理部件可以将附加字段加入每个数据记录中。这些附加字段可以称为流动单元。在一些实现中，每个流动单元标识通过数据流图的一段路径。该流动单元可以从数据记录中除去和存储起来供以后分析用。

[0030] 在一些实现中，数据流图工具化引擎本身可以包括接受数据流图作为输入和产生修改的数据流图的工具化数据流图。

[0031] 如过程箭头208所表示，可以将修改的数据流图提供给数据流图运行时环境210，数据流图运行时环境210可以是运行在一台计算机或多台计算机上的一个或多个进程。如过程箭头212所表示，将来自数据源214的数据记录提供给数据流图运行时环境210。数据流图运行时环境210使用修改的数据流图处理数据记录。

[0032] 如过程箭头216所表示，修改的数据流图将流动单元存储在流动单元中心库218中。流动单元中心库218可以是，例如，关系数据库或文件系统上的文件。

[0033] 如过程箭头220所表示，流动单元分析引擎222分析存储的流动单元。从存储的流动单元中，流动单元分析引擎222随着通过数据流图得到处理，可以确定每个数据记录走过的各种路径。流动单元分析引擎222可以确定至少一个数据记录走过的通过数据流图的所有不同路径。

[0034] 流动单元分析引擎222还可以确定记录相关性。在一些情况下，输出记录依赖于多个输入记录。例如，数据流图可以计算顾客在一年的过程中所下的订单的总值。每个订单代表一个单独输入记录，这些记录的总和产生单个 输出记录。流动单元分析引擎222可以确定用在输出记录的创建中的每个输入数据记录和每个中间数据。

[0035] 如过程箭头224所表示，流动单元分析引擎222可以存储描述通过修改的数据流图产生的数据记录的数据概要222。

[0036] 如过程箭头228所表示，流动单元分析引擎222还可以接受来自数据源214的数据记录。流动单元分析引擎222可以使用流动单元来确定数据记录的代表性样本。在一些实现中，代表性样本是这样确定的，使至少一个样本数据及其处理后上代记录在数据流图中走过每条不同路径。在一些实现中，代表性样本是这样确定的，使子集中的每个数据记录走过

相同路径。

[0037] 流动单元分析引擎222可以确定数据记录的子集，以便通过数据流图的记录的子集的流动覆盖区与处理来自数据源214的数据记录的整个集合时通过数据流图产生的流动覆盖区相比保持一致。例如，数据流图可以包括根据邮政编码聚集交易记录的数值的部件。当选择具有特定ZIP(邮政)编码的记录时，流动单元分析引擎222选择与ZIP编码相对应的所有数据记录。因此，无论处理数据记录的子集还是处理数据源214中的所有数据记录，那个邮政编码的聚集值都保持一致。在一些实现中，流动单元分析引擎222可以确定保持通过数据流图的流动覆盖区的分布的数据子集。例如，如果在处理数据记录的整个集合期间拒绝了10%的顾客，则当处理数据记录的子集时，也将拒绝10%的顾客。

[0038] 如过程箭头230所表示，流动单元分析引擎222可以将数据记录的子集存储在样本数据记录中心库232中。

[0039] 在一种实现中，暂停工具化数据流图的执行。在暂停状态下，通过让用户首先选择数据流图中的部件来选择一个部件中的数据记录。系统200向用户显示在暂停状态下停留在部件中的记录集合，以便用户接着可以二次选择用于观察的记录。当使用与所选记录相联系的流动单元数据时，系统200可以确定所有前记录(输入记录、和图形中直到包含所选记录的部件的部件产生的那些中间记录两者)的集合。当使用这个集合的输入记录时，系统200现在可以创建一批只包含这些所选记录的输入数据子集，然后利用这些数据子集重新启动该图形。该图形现在可以逐步调试所选记录的执行，以便使用户能够观察该图形直到用户首先选择的部件的执行的行为。

[0040] 图3例示了使用流动单元跟踪通过所修改数据流图的记录的例子。数据 源“ci”302包括两个数据记录，即，第一数据记录330和第二数据记录332。随着数据记录从数据源“ci”302提供给数据流图，将流动单元附在每个数据记录上。将流动单元304附在数据记录330上，将流动单元306附在数据记录332上。

[0041] 在一些实现中，该流动单元包括与该流动单元相联系的部件标识符、与该部件相联系的群、和序号。该群可以是，例如，提供数据记录的端口的指示。在一些实现中，该流动单元可以包括带有部件标识符、群标识符、和序号的格式化字符串(例如，“ci.a.1”、“r1.a.1”)。每个部件标识符能够唯一地标识数据流图中的部件。例如，流动单元304包括字符串“ci.a.1”，其中“ci”指示流动单元与数据源“ci”302相联系，群“a”指示在与字母“a”318相联系的端口上提供数据记录，以及序号“1”指示数据记录是在与字母“a”318相联系的端口上从数据源“ci”302提供的第一数据记录。

[0042] 类似地，流动单元306包括字符串“ci.a.2”，其中“ci”指示流动单元与数据源“ci”302相联系，群“a”指示在与字母“a”318相联系的端口上提供数据记录，以及序号“2”指示数据记录是在与字母“a”318相联系的端口上从数据源“ci”302提供的第二数据记录。在一些实现中，可以将部件和端口与数字、字母、字符串、或任何其他标识符相联系。

[0043] 在本例中，第一数据记录330和第二数据记录332两者都由过滤部件“r1”308处理。过滤部件“r1”308在第一端口320上提供第一数据记录330和在第一端口322上提供第一数据记录332。将新流动单元310提供给第一数据记录330。该新流动单元包括字符串“r1.a.1”，其中“r1”指示流动单元与过滤部件“r1”308相联系，“a”指示在“a”端口320上提供数据记录，以及“1”指示数据记录是在过滤部件“r1”308的“a”端口320上提供的第一数据

记录。

[0044] 过滤部件“r1”308处理在“b”端口322上提供第二数据记录332。与第一数据记录330类似，将新流动单元312提供给第二数据记录。该新流动单元312包括字符串“r1.b.1”，其中“r1”指示流动单元与过滤部件“r1”308相联系，“b”指示在“b”端口322上提供数据记录，以及“1”指示数据记录是在过滤部件“r1”308的“b”端口322上提供的第一数据记录。

[0045] 不是每个部件都向数据记录提供新流动单元。在本例中，部件314、部件316、和部件334被认为是通过部件。在流动分析进程中忽略这些部件。这里，第一数据记录330保留它的流动单元310，第二数据记录332保留它的流动单元312。在一些实现中，不变更数据记录的路径的部件不提供新流动单元。在其他实现中，每个部件都向数据记录提供新流动单元。流动工具化引擎可以有选择地确定将图形中的哪些部件工具化，以便构建新流动单元并对其指定黑色或群以反映已经通过了特定部件和端口。

[0046] 在将数据记录存储在数据宿324中之前，除去流动单元。将除去的流动单元存储在流动单元数据存储326中。在本例中，数据宿324存储第一数据记录330和第二数据记录332。流动单元数据存储326存储流动单元312和流动单元310。

[0047] 图4例示了消耗流动单元的例子。标识部件、端口、和序号不一定足以唯一标识通过数据流图的整条路径。为了标识整条路径，可以组合流动单元。例如，参照图4，数据流图包括部件“z1”402、部件“z2”404、和部件“r7”406。在本例中，数据记录412是要通过部件“z1”402的“a”端口提供的第四数据记录。将包括字符串“z1.a.4”的流动单元410提供给数据记录412。在数据记录412经过部件“r7”406处理之后，将新流动单元414提供给该数据记录。该新流动单元指示该数据记录是在部件“r7”406的“a”端口408上提供的第六数据记录(“r7.a.6”)。在没有更多的情况下，流动单元414不指示数据记录412由部件“z1”402提供还是由部件“z2”404提供。为了保留整条路径，将流动单元410并入流动单元414中(或被流动单元414消耗掉)。在一些实现中，流动单元消耗以前与数据记录或与经过处理产生数据记录的一个或多个数据记录相联系的一组其他流动单元。

[0048] 在其他实现中，流动单元保留对以前与数据记录或处理后数据记录相联系的流动单元的引用，或以前与数据记录或处理后数据记录相联系的流动单元的副本。随着每个流动单元被取代，将旧流动单元存储在流动单元中心库(例如，图2的流动单元中心库218中)。新创建的流动单元包含对存储在流动单元中心库中的被取代流动单元的引用。图5例示了使用流动单元跟踪通过数据流图的路径的例子。可以将流动单元用于跟踪通过数据流图的复杂路径。这些复杂流动可以使用描述性字符串来描述，以便每个数据记录及其沿着通过数据流图的相同路径的上代具有相同字符串。通过将已执行流动路径的集合与通过数据流图的所有可能流动流径的集合相比较，可以确定用于生成已执行流动路径的集合的数据记录是否足以测试数据流图中的所有路径(即，已执行流动路径是否覆盖所有可能流动路径)。

[0049] 将每个流动单元与一个数据记录相联系，由于流动单元还引用其他“消耗掉”流动单元，所以可以标识经过处理产生每个流动单元的记录的整个集合。通过选择所生成流动单元的子集，可以选择与所选流动单元相对应的数据记录的子集。例如，可以将数据流图500用于剖析顾客人口统计。在本例中，部件“ac”502提供代表一组顾客“a”的数据记录流(简称为“a”顾客)。划分部件“f1”504将数据记录划分成两个集合，每个集合在不同输出端

口上提供。例如，划分部件“f1”504根据ZIP编码来划分“a”顾客。

[0050] 部件“at”510提供代表“a”顾客完成的一组交易的流动数据记录（简称为“a”交易）。

[0051] 联接部件512联接顾客数据记录和顾客交易产生“a”顾客的组合数据记录流。不能与交易数据记录联接的顾客数据记录和不能与顾客数据记录联接的交易数据记录在通向出口点514的单独输出端口上提供。一般说来，将每个顾客交易数据记录与完成交易的顾客相联系。在本例中，顾客可以用来自划分部件“f1”504的“a”端口的顾客数据记录或来自划分部件“f1”504的“b”端口的顾客数据记录来表示，但不是两者。

[0052] 部件“bc”516提供代表一组顾客“b”的数据记录流（简称为“b”顾客）。划分部件“f2”518将数据记录划分成两个集合，每个集合在不同输出端口上提供。例如，划分部件“f2”518根据ZIP编码来划分“b”顾客。

[0053] 部件“bt”524提供代表“b”顾客完成的一组交易的流动数据记录（简称为“b”交易）。

[0054] 联接部件“j2”526联接“b”顾客数据记录和“b”顾客交易产生“b”顾客的组合数据记录流。不能与交易数据记录联接的顾客数据记录和不能与顾客数据记录联接的交易数据记录在通向出口点530的单独输出端口上提供。一般说来，将每个顾客交易数据记录与完成交易的顾客相联系。在本例中，顾客可以用来自划分部件“f2”518的“a”端口的顾客数据记录或来自划分部件“f2”518的“b”端口的顾客数据记录来表示，但不是两者。

[0055] 部件“ci”532提供代表一般顾客信息的数据记录流（简称为顾客信息）。部件“r1”534重新格式化提供顾客信息的数据记录。不能重新格式化的数据记录在通往出口点536的端口上提供。可以重新格式化的数据记录在第二端口上提供。

[0056] 联接部件“j3”538将重新格式化的顾客信息数据记录与来自联接部件“j1”512和数据记录和来自联接部件“j2”526的数据记录组合。不能联接的数据记录在通往出口点540的输出端口上提供。联接的数据记录在输出端口上提供，流向汇总部件“ru”542。一般说来，部件“j3”538将重新格式化的顾客信息数据与来自划分部件“j1”512的“a”端口的数据记录或与来自划分部件“j2”526的“a”端口的数据记录组合。

[0057] 汇总部件“ru”542根据一些准则，例如，根据邮政编码来聚集顾客交易记录。聚集的记录在通向出口点544的端口上提供。

[0058] 将数据流图500修改成使用流动单元来跟踪数据记录的流动。例如，使用数据流图工具化引擎（例如，显示在图2中的数据流图工具化引擎206）。当使用字符串表示时，流动单元可以与可以组合起来描述流动的各个数据记录无关地描述数据记录通过数据流图的复杂流动。例如，由部件“ac”提供的数据记录可以用指示数据记录源自部件“ac”502的“a”端口的流动单元字符串“ac.a”来标记。

[0059] 一旦过滤部件“f1”504过滤了数据记录，就可以用指示数据记录是在过滤部件“f1”504的“b”端口上提供的流动单元字符串“f1.b”标记数据记录。在一些实现中，流动单元可以包括标识数据记录走过的路径的每个部分的历史字符串。例如，可以用历史字符串“(ac.a)f1.b”标记数据记录。括号可以用于指示f1.b流动单元消耗了ac.a流动单元。

[0060] 类似地，部件“at”510提供的数据记录用带有指示数据记录是在部件“at”510的端口上提供的历史字符串“at.a”的流动单元来标记。

[0061] 联接部件“j1”512将来自过滤部件“f1”504的一个顾客数据记录与来自部件“at”510的交易记录组合。组合的数据记录可以用带有历史字符串“((ac.a)f1.b,at.a)j1.a”的流动单元来标记，历史字符串“((ac.a)f1.b,at.a)j1.a”指示通过组合来自顾客记录的流动单元(“(ac.a)f1.b”)和来自交易记录的流动单元(“at.a”)创建了新的流动单元。逗号可以用于隔开组合在一起的多个流动单元。

[0062] 类似地，顾客信息数据记录由部件“ci”532提供，并用具有历史字符串“ci.a”的流动单元来标记。顾客信息数据记录由部件“r1”534重新格式化。重新格式化的顾客信息数据记录用具有历史字符串“(ci.a)r1.a”的流动单元来标记。

[0063] 联接部件“j3”538组合来自联接部件“j1”512的组合数据记录和来自部件“r1”534的重新格式化顾客信息数据记录。新组合记录可以用具有历史字符串“(((ac.a)f1.b,at.a)j1.a,(ci.a)r1.a)j3.a”的流动单元来标记。

[0064] 汇总部件“ru”542将源自联接部件“j3”538的多个记录组合成单个记录。历史字段中的星号“*”可以用于指示将来自相同流的多个记录组合在一起。例如，汇总记录可以用带有历史字符串“(*(((ac.a)f1.b,at.a)j1.a,(ci.a)r1.a)j3.a)ru.a”的流动单元来标记。这个记号指示在产生ru.a流动单元时消耗了多个“(((ac.a)f1.b,at.a)j1.a,(ci.a)r1.a)j3.a”流动单元。

[0065] 图6例示了识别导致输出数据记录的产生的输入数据记录的例子。一旦确定了所有执行流动路径历史，就可以将流动路径历史用于确定在所有或一些执行流动路径上提供全部或部分覆盖区的数据子集。例如，存储在关系数据库中的表格602包括从图5的数据流图500中产生的所有数据记录的执行流动路径。表格602的每一行对应于数据流图产生的输出数据记录。表格的一列包括描述取来产生输出记录的执行路径的流动单元历史。从输出记录中可以确定输入记录。每个输出记录对应于包括如上面参照图4所述的嵌套流动单元的流动单元。例如，如箭头606所表示，行604包含流动单元608的表示。流动单元608包括在创建流动单元时消耗的所有流动单元。产生与行604相对应的输出数据记录所需的输入数据记录可以通过检查流动单元树的“叶节点”来确定。那是在创建它们时未消耗任何其他流动单元的流动单元。

[0066] 在本例中，嵌套流动单元610指示在创建输出数据记录时使用了在部件“ci”的“a”端口上提供的第10数据记录。由于部件“ci”只包含单个端口(参见图5的部件“ci”532)，所以可以唯一地标识输入数据记录。

[0067] 类似地，嵌套流动单元612,614,616,618和620指示在创建输出数据记录时分别使用了在部件“at”的“a”端口上提供的第8数据记录、在部件“ac”的“a”端口上提供的第10数据记录、在部件“ci”的“a”端口上提供的第19数据记录、在部件“at”的“a”端口上提供的第21数据记录、和在部件“ac”的“a”端口上提供的第30数据记录。包括这些记录的数据的样本集合产生与表格602的行604相对应的输出数据记录。

[0068] 通过使用与输出数据记录的所选子集相联系的流动单元历史，可以选择产生这些输出记录的输入数据记录的子集。输入数据记录的这个子集可以不变更程序的行为地用于测试或分析。例如。如果测试集合未包括在部件“at” 的“a”端口上提供的第21数据记录，则该图形的执行将不产生相同输出记录。

[0069] 图7例示了修改数据流图的数据源部件以便将流动单元加入数据记录中的例子。

一般说来，数据流图可以被工具化成将流动单元加入数据记录中，将流动单元组合在一起提供到数据流图的执行路径的地图，以及在数据记录从数据流图中退出之前除去流动单元。

[0070] 修改数据流图700提供记录谱系可以包括将重新格式化每个数据记录以包括流动单元的部件加在每个数据源的后面。在一些实现中，用包含原始输入数据集合部件的副本和重新格式化每个数据记录以包括流动单元的部件的子图(或嵌套图)取代每个数据源。例如，将数据流图700修改成数据源702将数据记录提供给流动单元生成器部件704。

[0071] 在一些实现中，流动单元生成器部件704将附加字段加入每个数据记录中，该附加字段是如上所述的流动单元。

[0072] 图8例示了修改有多个输出端口的部件以便将流动单元加入数据记录中的例子。数据流图700包括具有多个输出端口804，806的重新格式化部件802。在本例中，为每个端口加入单独流动单元生成器。端口804对应于流动单元生成器808，端口806对应于流动单元生成器810。在一些实现中，用子图取代带有多个输出端口的部件。例如，子图812包含原始部件的副本，以及在每个端口上将流动单元提供给数据记录。

[0073] 图9例示了修改数据宿以便处理流动单元的例子。数据流图被修改成加入了流动单元除去部件904。流动单元除去部件904从数据记录中除去流动单元，并将其存储在流动单元中心库906中。将没有流动单元的数据记录存储在数据宿902中。

[0074] 图10例示了跨过多个数据流图地使用流动单元的例子。例如，工具化数据流图A1004从数据存储1002中读取数据记录。如过程箭头1006所表示，数据流图A1004处理该记录并产生到数据存储1008的输出数据记录。如过程箭头1012所表示，将数据流图A1004产生的每个输出数据记录与存储在流动单元中心库1010中的流动单元相联系。

[0075] 如过程箭头1016所表示，工具化数据流图B从数据存储1008中读取数据记录，处理该记录，并将它们存储在数据存储1020中。如过程箭头1014所表示，取代为从数据存储1008中读取的每个数据记录创建新流动单元，工 具化数据流图B1018从流动单元中心库1010中读取与数据记录相联系的流动单元。

[0076] 如过程箭头1022所表示，工具化数据流图将流动单元存储在流动单元中心库1010中。从工具化数据流图B1018中产生的流动单元包括消耗的流动单元和标识通过数据流图A1004和数据流图B1018两者的数据记录及其上代数据记录的全部执行路径的历史。

[0077] 图11例示了流动分析的示范性过程。该过程可以在一台或多台计算设备，例如，图1的计算机112上实现。

[0078] 过程1100修改(1102)数据流图。该数据流图被修改成加入和从通过数据流图处理的数据记录中除去流动单元。每个流动单元标识通过数据流图的一段路径。

[0079] 过程1100根据流动单元识别(1104)执行路径。该流动单元可以用于识别一个数据记录和用于产生它的前记录走过的通过数据流图的路径。

[0080] 过程1100确定(1106)数据记录的子集。该数据记录的子集根据执行路径来确定，以便该子集中的至少一个数据记录走过通过数据流图的每条执行路径。

[0081] 描述在本说明书中的主题的实施例和操作可以在数字电路中，或在包括公开在本说明书中的结构和它们的结构等效物的计算机软件、固件或硬件中，或在它们的两种或更多种的组合体中实现。描述在本说明书中的主题的实施例可以实现成编码在计算机存储介

质上供数据处理装置执行或控制数据处理装置的操作的一个或多个计算机程序，即，计算机程序指令的一个或多个模块。可替代地，或另外，可以将程序指令编码在人工生成传播信号，例如，为编码信息而生成，发送给适当接收装置的机器生成电、光、或电磁信号上以便供数据处理装置执行。计算机存储介质可以是计算机可读存储设备、计算机可读存储基片、随机或串行访问存储阵列或设备、或它们的两种或更多种的组合体，或包括在它们当中。此外，虽然计算机存储介质不是传播信号，但计算机存储介质可以是编码在人工生成传播信号中的计算机程序指令的源头或目的地。计算机存储介质还可以是一个或多个单独物理部件或介质(例如，多个CD、盘、或其他存储设备)，或包括在它们当中。

[0082] 描述在本说明书中的操作可以实现成由数据处理装置对存储在一个或多个计算机可读存储设备上或从其他源头接收的数据进行的操作。

[0083] 术语“数据处理装置”包含处理数据的所有类型装置、设备、和机器，举例来说，包括可编程处理器、计算机、芯片上的系统、或上述的多个或组合体。该装置可以包括专用逻辑电路，例如，FPGA(现场可编程门阵列)或ASIC(专用集成电路)。除了硬件之外，该装置还可以包括为所涉及的计算机程序创造执行环境的代码，例如，构建处理器固件、协议栈、数据库管理系统、操作系统、跨平台运行时环境、虚拟机、或它们的两种或更多种组合体的代码。该装置和执行环境可以实现像万维网服务、分布式计算和网格计算基础设施那样的各种不同计算模型基础设施。

[0084] 计算机程序(也称为程序、软件、应用软件、脚本、或代码)可以用包括编译或解释语言、或说明或过程语言的任何形式编程语言来编写，并且可以以包括部件、子例程、对象、或适合用在计算环境中的其他单元作为独立程序或作为模块的任何形式部署。计算机程序可以，但无需，对应于文件系统中的文件。可以将程序存储在保存其他程序或数据(例如，存储在标记语言文档中的一个或多个脚本)的文件的一部分中，存储在专用于所涉及的程序的单个文件中，或存储在多个协作文件(例如，存储一个或多个模块、子程序和部分代码的文件)中。可以将计算机程序部署成在单台计算机上或在位于单个地点或分布在多个地点上和通过通信网络互连的多台计算机上执行。

[0085] 描述在本说明书中的进程和逻辑流程可以由执行一个或多个计算机程序以便通过操作输入数据和生成输出来执行动作的一台或多台可编程处理器来执行。该进程和逻辑流程也可以由专用逻辑电路，例如，FPGA(现场可编程门阵列)、或ASIC(专用集成电路)来执行，以及也可以将装置实现成专用逻辑电路，例如，FPGA(现场可编程门阵列)、或ASIC(专用集成电路)。

[0086] 举例来说，适合执行计算机程序的处理器包括通用和专用微处理器两者、和任何类型数字计算机的任何一个或多个处理器。一般说来，处理器从只读存储器或随机访问存储器或两者接收指令和数据。计算机的基本元件是依照指令执行动作的处理器、和存储指令和数据的一个或多个存储设备。一般说来，计算机还包括存储数据的一个或多个海量存储设备，例如，磁盘、磁光盘、或光盘，或可操作地耦合成从它们那里接收数据或将数据传送给它们，或两者。但是，计算机无需含有这样的设备。此外，可以将计算机嵌入另一个设备，例如，移动电话、个人数字助理(PDA)、移动音频或视频播放器、游戏机、全球定位系统(GPS)接收器、或便携式存储设备(例如，通用串行总线(USB)闪存驱动器)等中。适合存储计算机程序指令和数据的设备包括所有形式的非易失性存储器、介质和存储设备，举例来说，

包括半导体存储设备，例如，可擦除可编程只读存储器(EPROM)、电可擦除可编程只读存储器(EEPROM)、和闪速存储设备；磁盘，例如，内部硬盘或可换式盘；磁光盘；以及CD-ROM、和DVD-ROM盘。处理器和存储器可以通过专用逻辑电路来补充，或并入专用逻辑电路中。

[0087] 为了提供与用户的交互，描述在本说明书中的主题的实施例可以在含有向用户显示信息的显示设备，例如，CRT(阴极射线管)、或LCD(液晶显示器)监视器以及用户可以向计算机提供输入的键盘和定位设备，例如，鼠标或跟踪的计算机上实现。也可以将其他类型的设备用于提供与用户的交互；例如，提供给用户的反馈可以是任何形式的感觉反馈，例如，视觉反馈、听觉反馈或触觉反馈；以及来自用户的输入可以以任何形式接收，包括声音、语音或触觉输入。另外，计算机可以通过将文档发送给用户使用的设备或从用户使用的设备接收文档，例如，通过响应从万维网浏览器接收的请求，将网页发送给用户客户设备上的万维网浏览器与用户交互。

[0088] 描述在本说明书中的主题的实施例可以在包括后端部件，例如，作为数据服务器，包括中间部件，例如，应用服务器，包括前端部件，例如，具有用户可以与描述在本说明书中的主题的实现交互的图形用户界面或万维网浏览器的客户计算机，或包括两个或更多个这样的后端、中间、或前端部件的任何组合体的计算系统中实现。系统的部件可以通过数字数据通信的任何形式或介质，例如，通信网络互连。通信网络的例子包括局域网(LAN)、广域网(WAN)、互联网(例如，因特网)、和对等网络(例如，对点对对等网络)。

[0089] 计算系统可以包括客户机和服务器。客户机和服务器一般相互远离，通常通过通信网络交互。客户机和服务器的关系通过运行在各自计算机上和相互具有客户机-服务器关系的计算机程序建立起来。在一些实施例中，服务器将数据(例如，HTML页面)发送给客户设备(例如，为了向与客户设备交互的用户显示数据和从该用户接收用户输入的目的)。可以在服务器上从客户设备接收在客户设备上生成的数据(例如，用户交互的结果)。

[0090] 虽然本说明书包含许多具体实现细节，但这些不应该被理解为限制任何发明的范围或可能要求保护的范围，而是作为对特定发明的特定实施例特有 的特征的描述。在单独实施例的背景下描述在本说明书中的某些特征也可以在单个实施例中以组合形式实现。相反，在单个实施例的背景下描述的各种特征也可以在多个实施例中单独地或以任何适当分组合的形式实现。此外，尽管上面可能将一些特征描述成以某种组合形式起作用和甚至最初要求这样，但所要求组合当中的一种或多种特征在一些情况下可以从该组合中分割出来，以及所要求组合可以针对分组合或分组合的变种。

[0091] 类似地，虽然在附图中按特定次序描述这些操作，但不应该理解为要求按所示的特定次序或顺序地执行这样的操作，或执行所有例示的操作来获得所希望的结果。在某些情况下，多任务和并行处理可能是有利的。此外，在上述的实施例中各种系统部件的分离不应该理解为在所有实施例中都要求这样的分离，而应该理解为所述程序部件和系统一般可以一起合并在单个软件产品中或打包成多个软件产品。

[0092] 因此，我们已经描述了主题的特定实施例。其他实施例在所附权利要求书的范围之内。在一些情况下，列举在权利要求书中的动作可以按不同次序执行，但仍然可以获得所希望的结果。另外，描述在附图中的过程未必要求所示的特定次序或顺序来获得所希望的结果。在某些实现中，多任务和并行处理可能是有利的。

[0093] 上述的流动分析手段可以使用运行在计算机上的软件来实现。例如，该软件形成

运行在一个或多个编程或可编程计算机系统(可以具有像分布式、客户机/服务器、或网格那样的各种架构)上的一个或多个计算机程序中的过程,每个编程或可编程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、和至少一个输出设备或端口。该软件可以形成,例如,提供与数据流图的设计和配置有关的其他服务的较大程序的一个或多个模块。该图形的节点和元件可以实现成存储在计算机可读介质中的数据结构或遵从存储在数据中心库中的数据模型的其他有组织数据。

[0094] 软件可以在像CD-ROM那样,通用或专用可编程计算机可读的存储介质上提供,或在网络的通信介质上输送(编码在传播信号中)给执行它的计算机的存储介质。所有功能都可以在专用计算机上执行,或使用像协处理器那样的专用硬件来执行。软件可以实现成不同计算机执行软件规定的计算的不同部分的分布式。每个这样的计算机程序优选地存储在通用或专用可编程计 算机可读的存储介质或设备(例如,固态存储器或介质、或磁或光介质)上,或下载到这样的存储介质或设备,以便当计算机系统读取该存储介质或设备执行描述在其中的过程时配置和操作计算机。本发明系统也可以被认为实现成配有计算机程序的计算机可读存储介质,其中如此配置的存储介质使计算机系统以特定和预定方式操作,以执行描述在其中的功能。

[0095] 上面已经描述了本发明的许多实施例。不过,应该明白,可以不偏离本发明的精神和范围地作出各种修改。例如,上述的一些步骤可以是与次序无关的,因此可以按与所述不同的次序执行。

[0096] 要明白的是,前面的描述旨在例示本发明而不是限制本发明的范围,本发明的范围由所附权利要求书的范围限定。例如,上述的许多功能步骤可以基本上不影响整个处理地按不同次序执行。其他实施例在所附权利要求的范围之内。

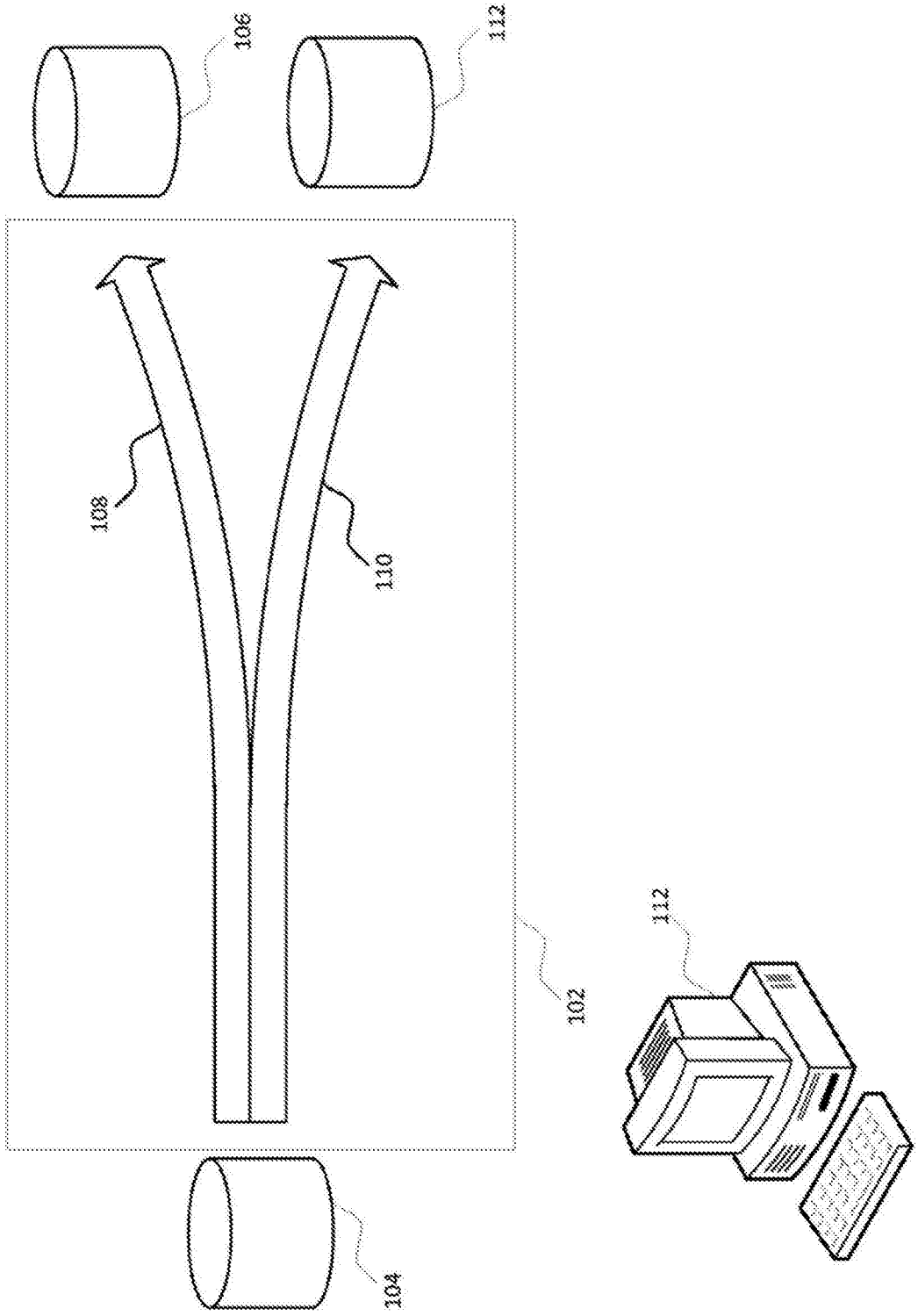


图1

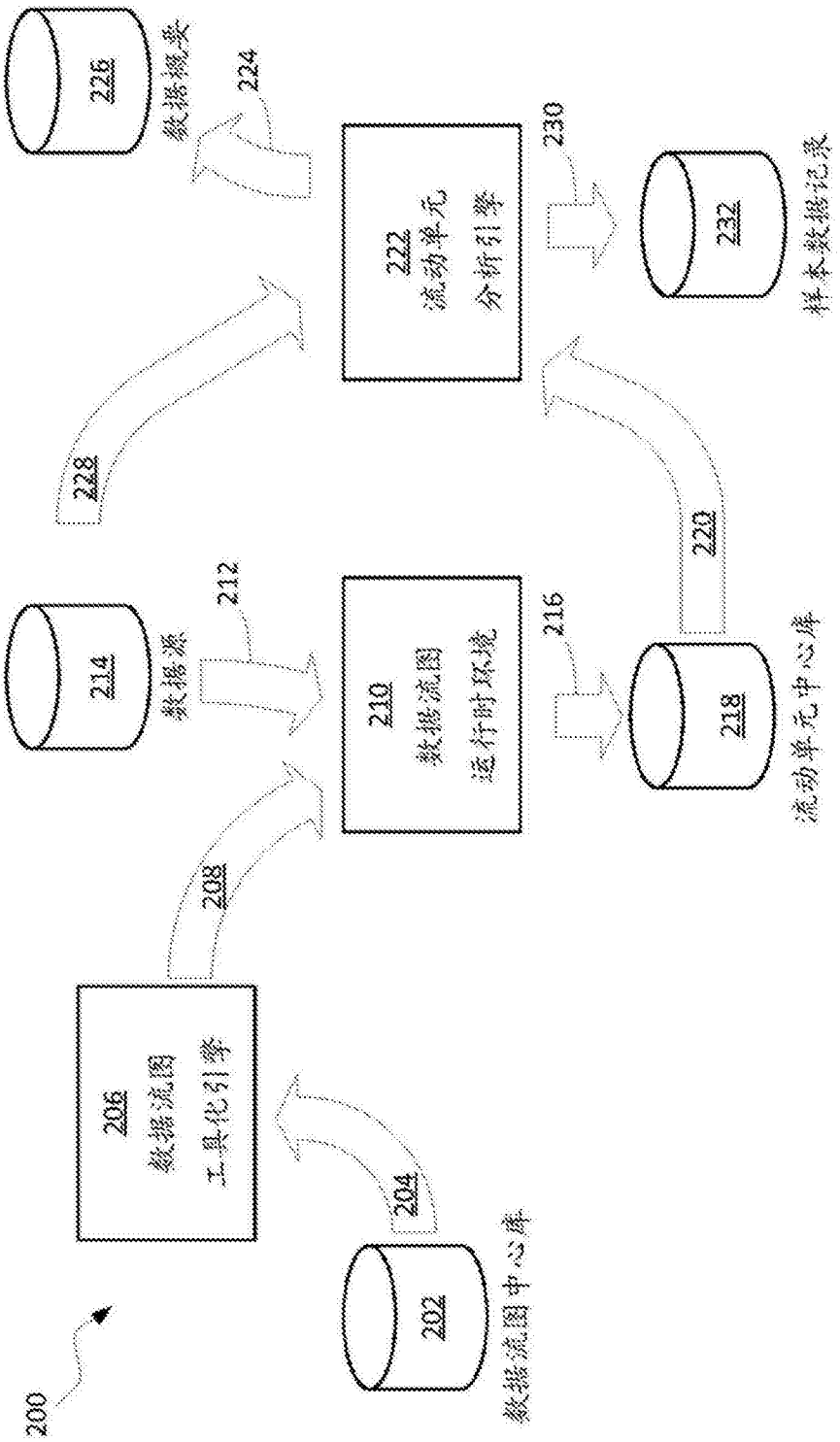

200
202
数据流图中心库
204
206
数据流图
工具化引擎
208
210
数据流图
运行时环境
212
214
数据源
216
218
流动单元中心库
220
222
流动单元
分析引擎
224
226
数据规类
228
230
232
样本数据记录

图2

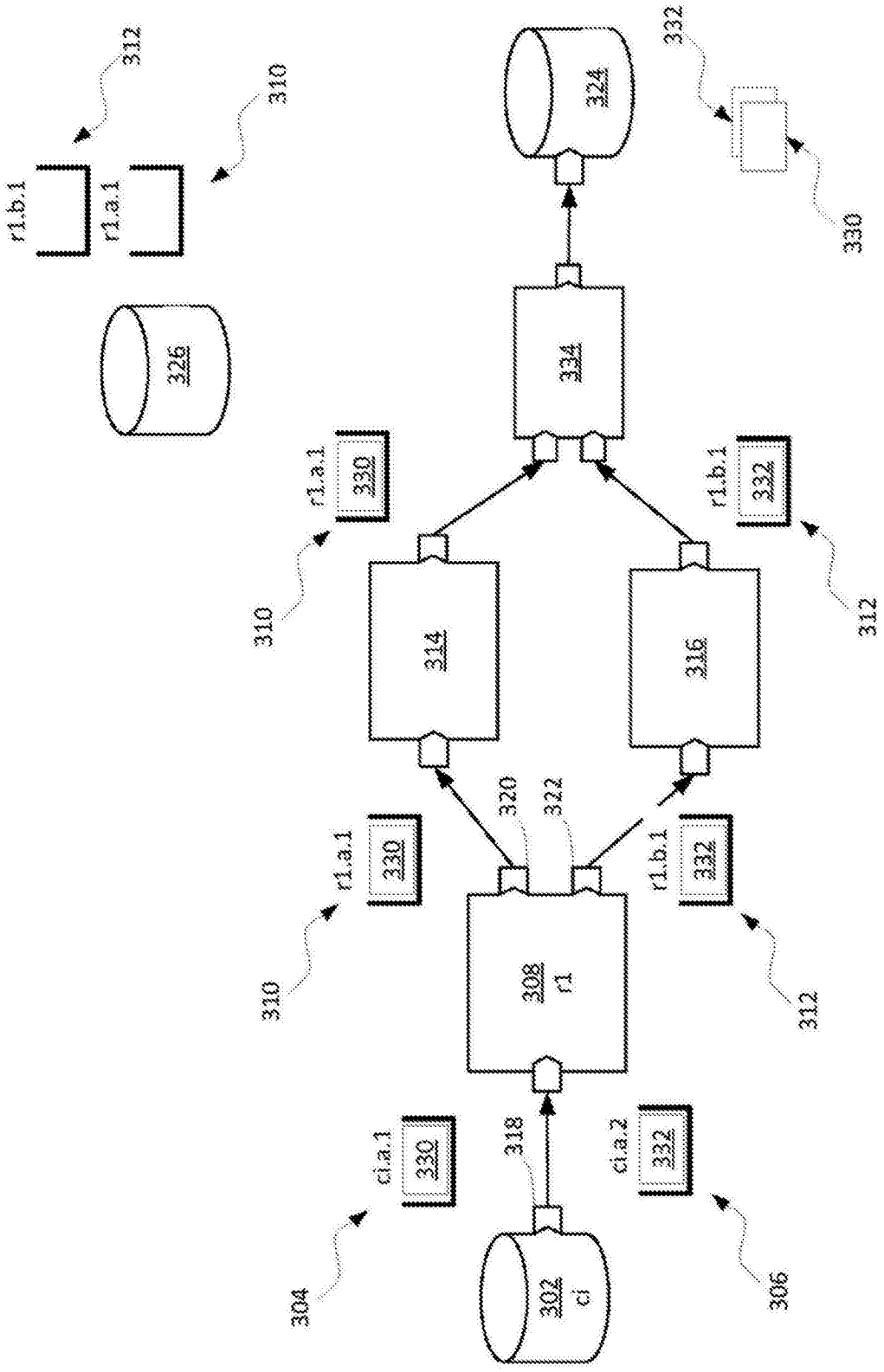

312
310
r1.b.1
r1.a.1
326
324
332
330
334
r1.a.1
330
310
r1.b.1
332
312
314
316
320
322
r1.a.1
330
310
r1.b.1
332
312
308
r1
c1.a.1
330
318
c1.a.2
332
304
306
302
ci

图3

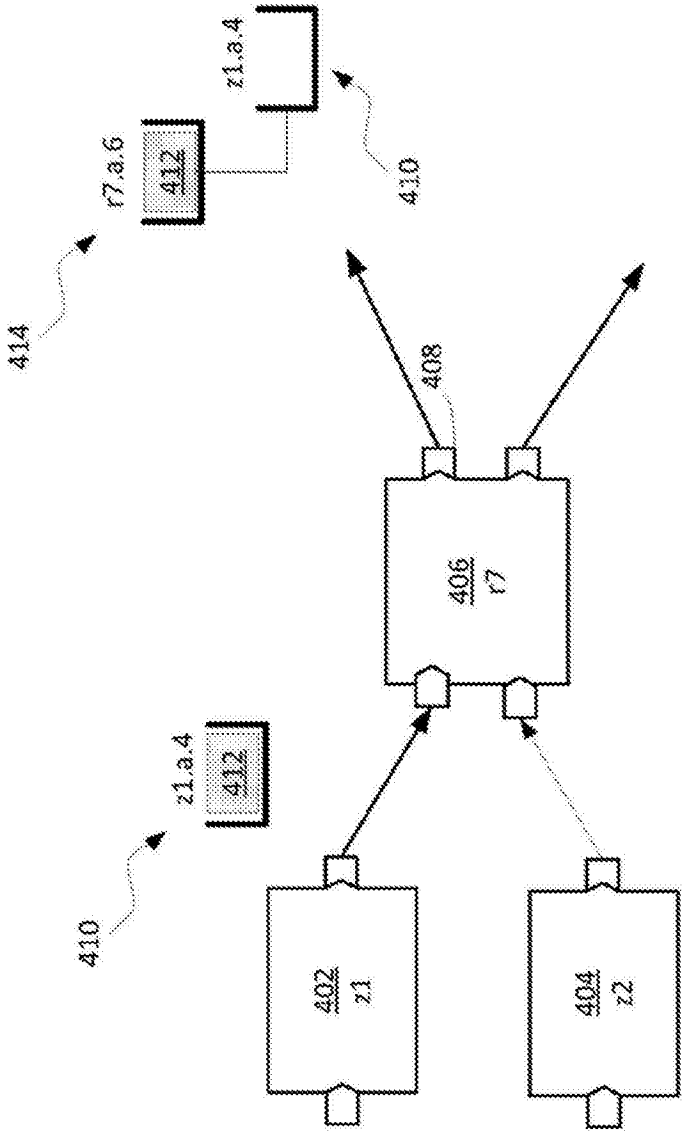


图4

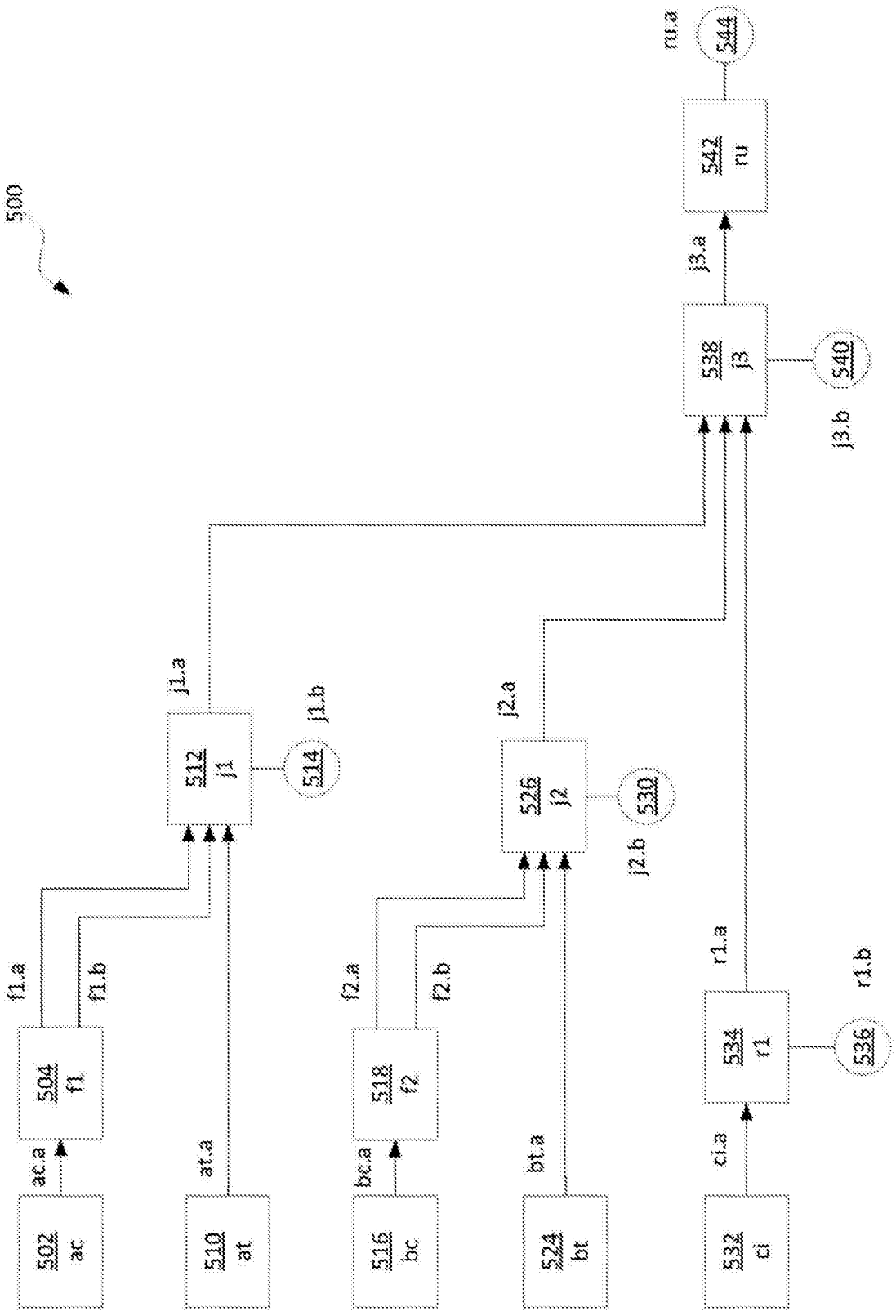

500
502 ac
504 f1
510 at
512 j1
514
516 bc
518 f2
524 bt
526 j2
530
532 ci
534 r1
536
538 j3
540
542 ru
544
ac.a
at.a
bc.a
bt.a
ci.a
f1.a
f1.b
f2.a
f2.b
j1.a
j1.b
j2.a
j2.b
j3.a
j3.b
r1.a
r1.b
ru.a

图5

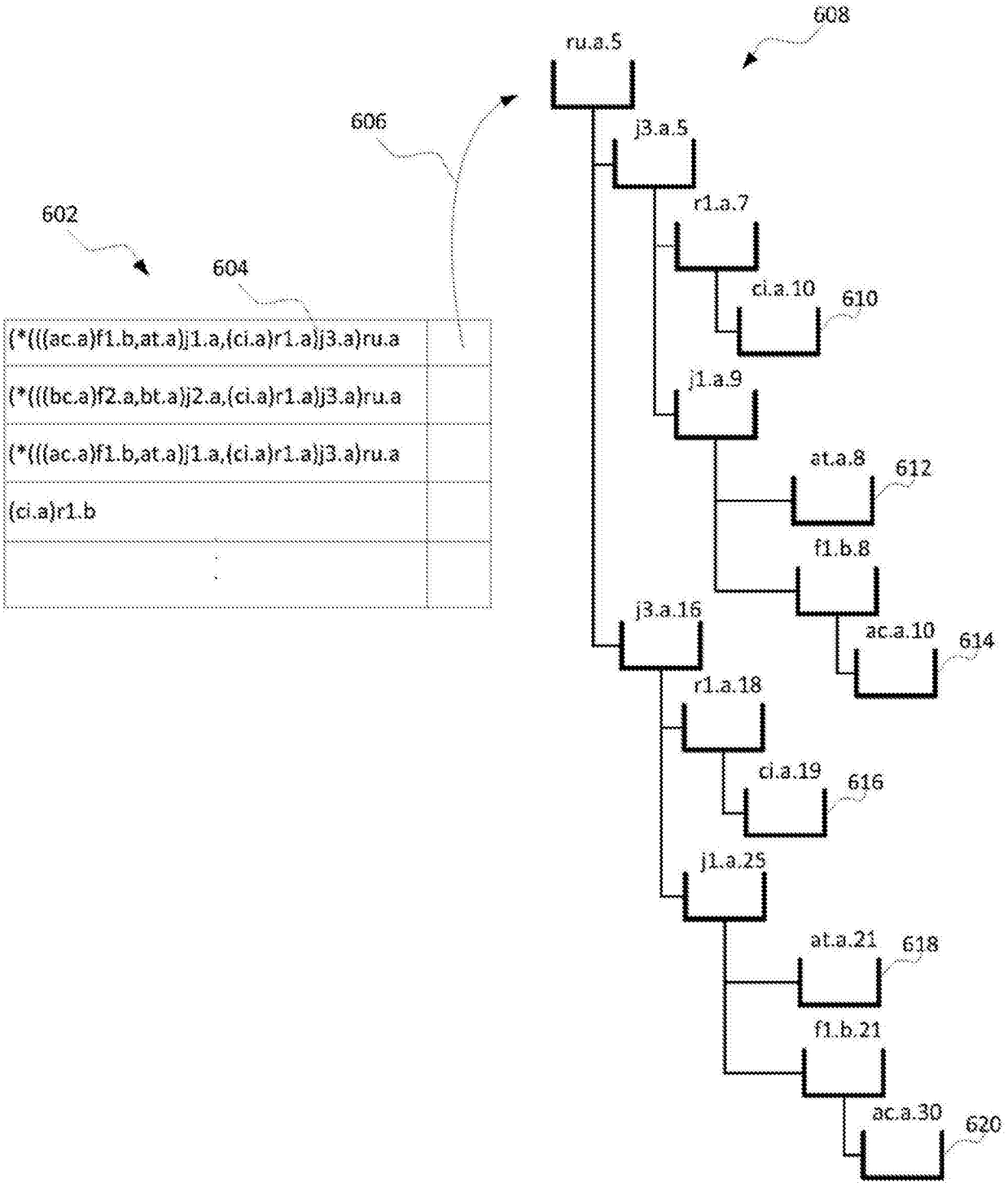

602
604
606
608
(*(((ac.a)f1.b,at.a)j1.a,(ci.a)r1.a)j3.a)ru.a
(*(((bc.a)f2.a,bt.a)j2.a,(ci.a)r1.a)j3.a)ru.a
(*(((ac.a)f1.b,at.a)j1.a,(ci.a)r1.a)j3.a)ru.a
(ci.a)r1.b
ru.a.5
j3.a.5
r1.a.7
ci.a.10
610
j1.a.9
at.a.8
612
f1.b.8
j3.a.16
ac.a.10
614
r1.a.18
ci.a.19
616
j1.a.25
at.a.21
618
f1.b.21
ac.a.30
620

图6

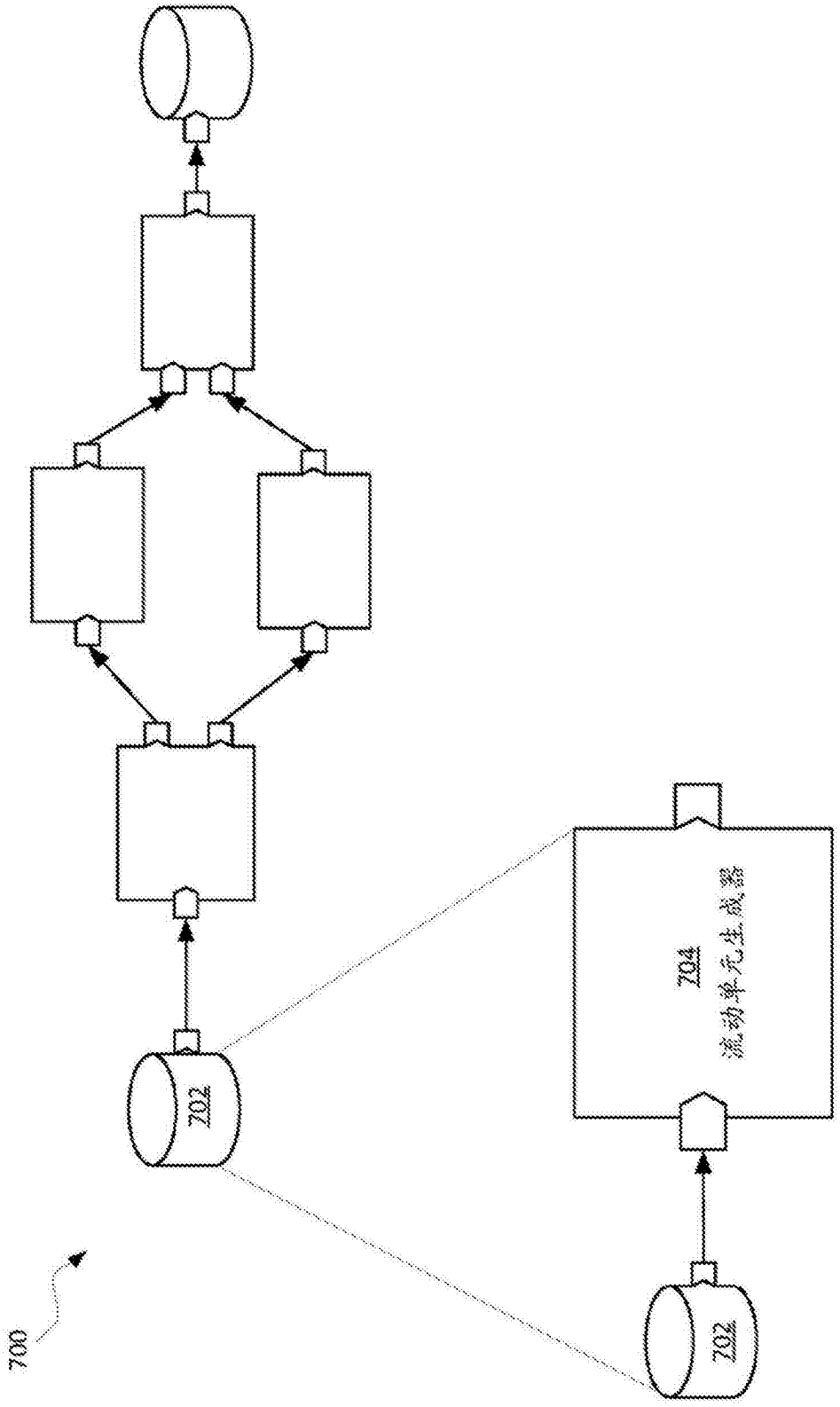


图7

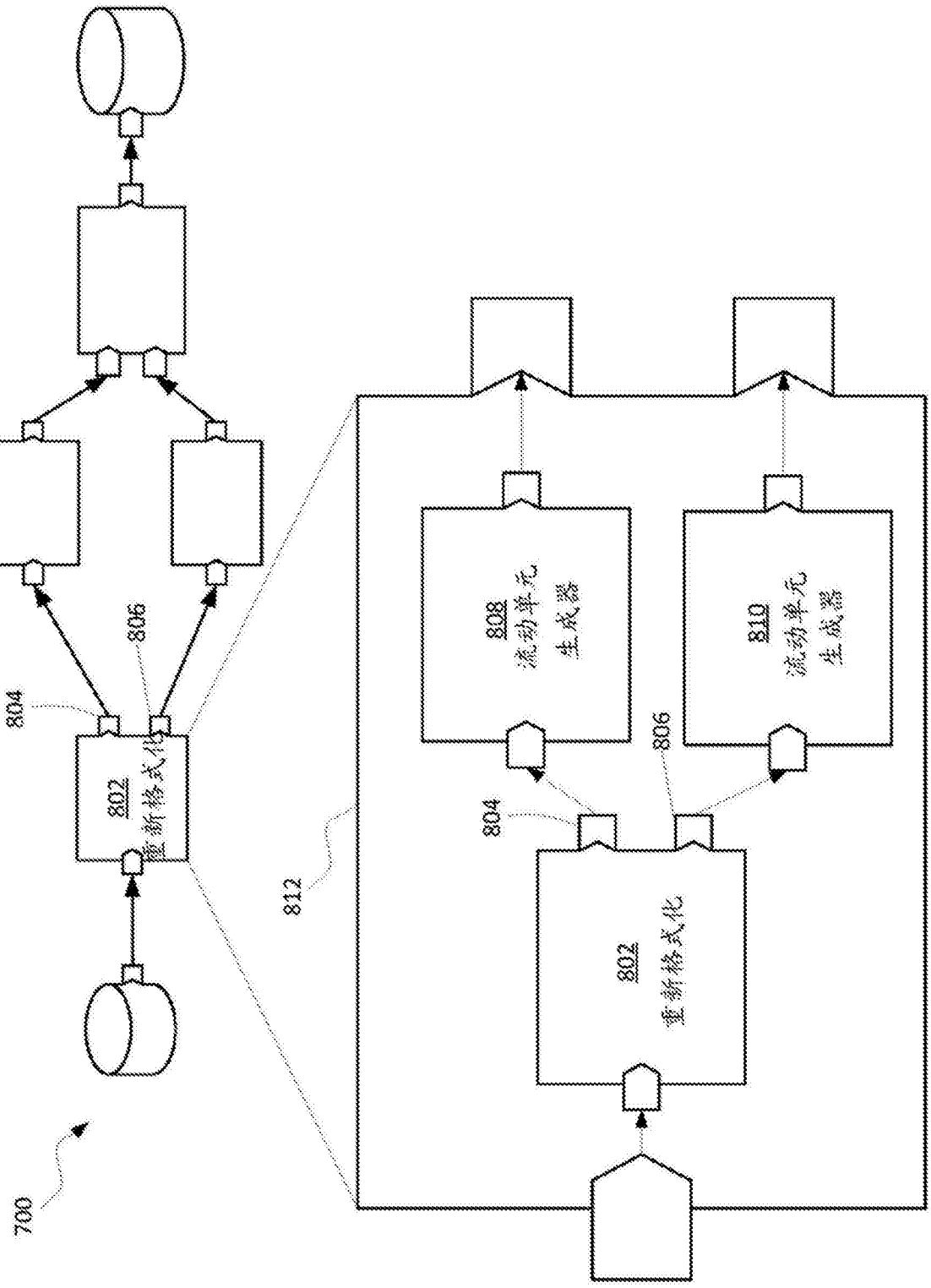

700
804
806
802
重新格式化
812
808
流动单元
生成器
810
流动单元
生成器
804
806
802
重新格式化

图8

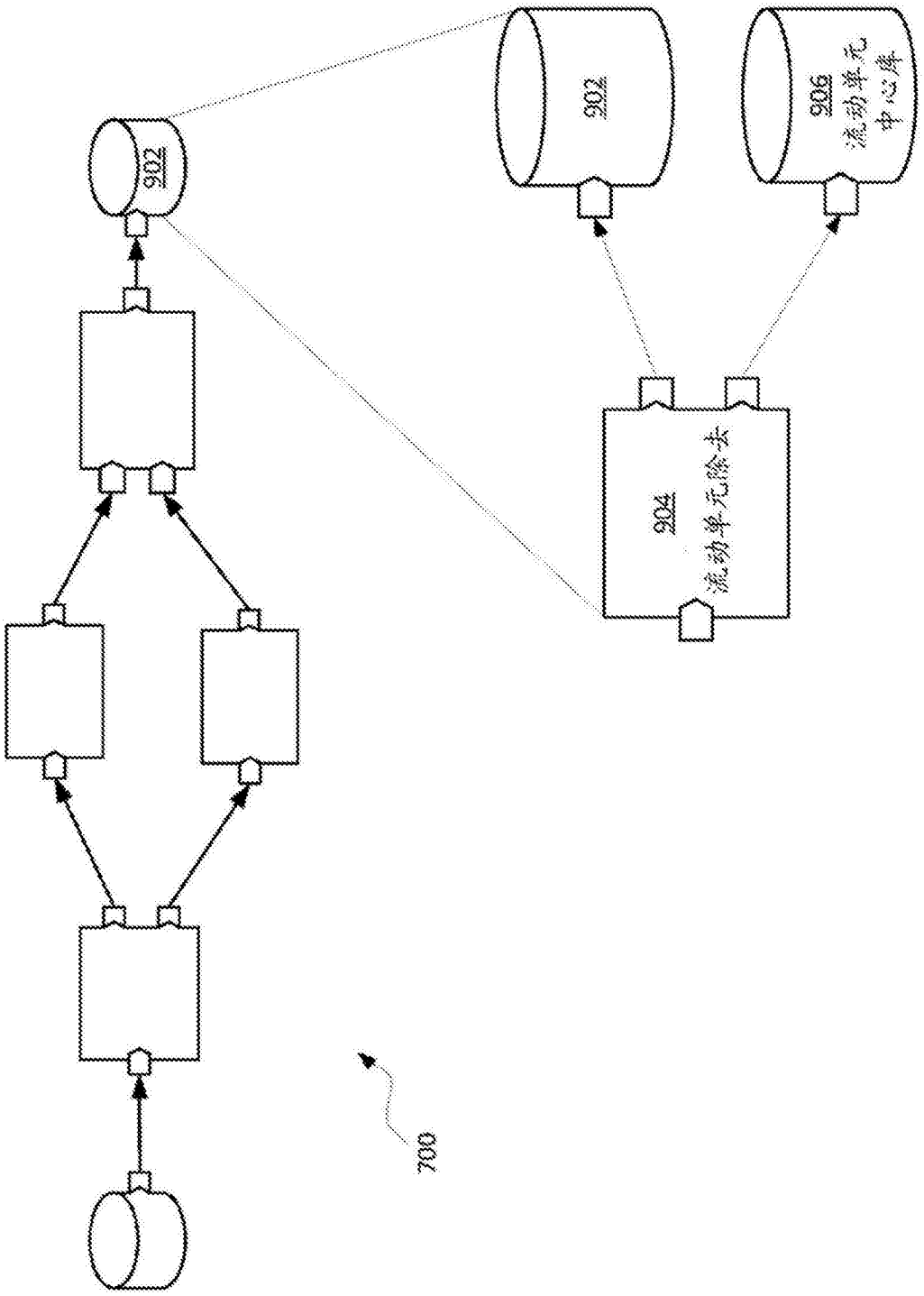


图9

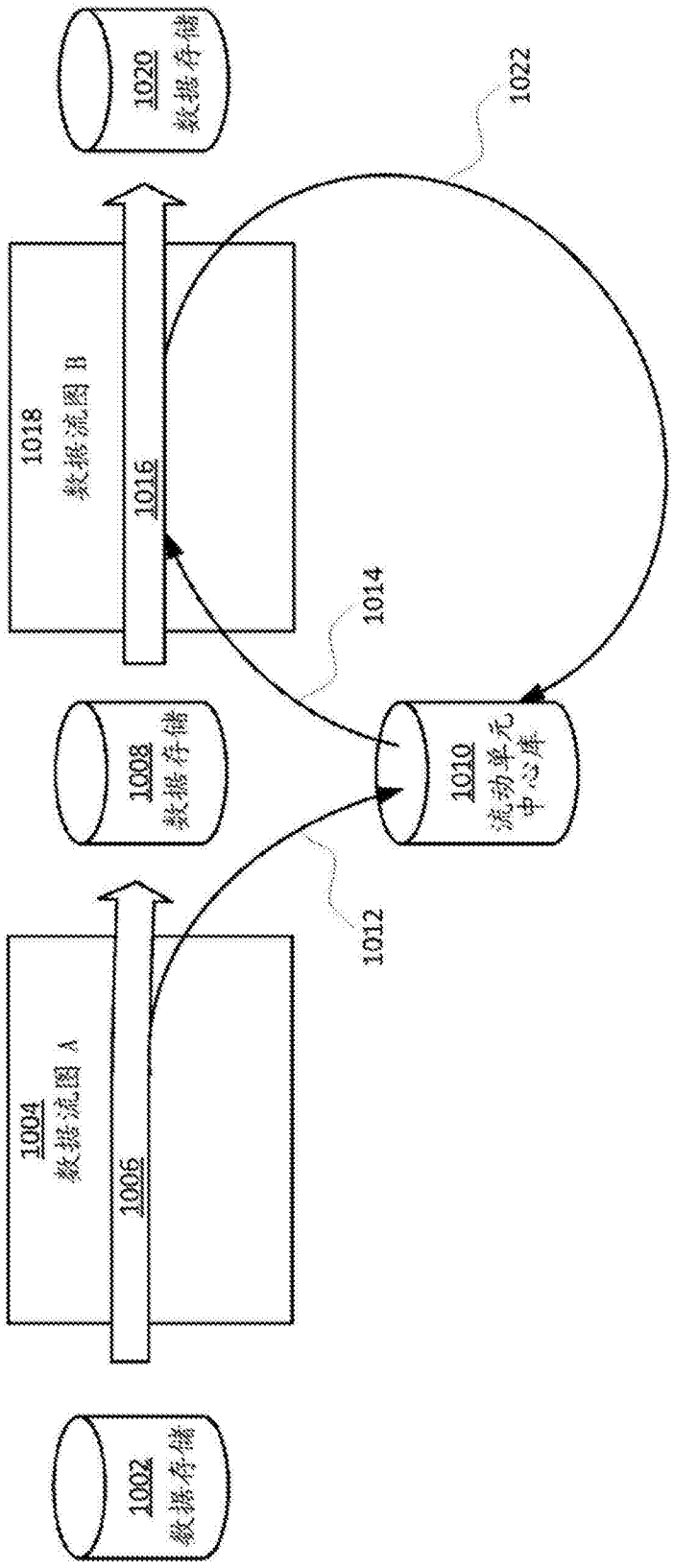

1020
数据存储
1022
1018
数据流图 B
1016
1014
1008
数据存储
1010
流动单元
中心库
1012
1004
数据流图 A
1006
1002
数据存储

图10

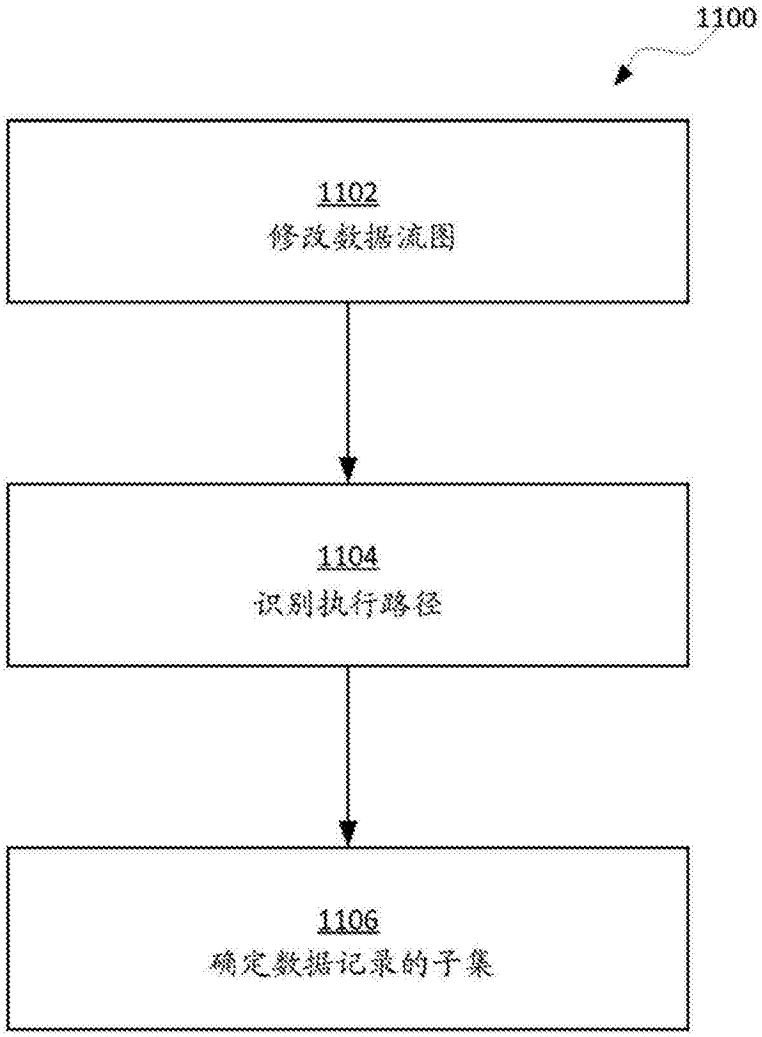


图11