(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 107430611 B
(45)授权公告日 2021.12.31

(21)申请号 201680009992.1

(22)申请日 2016.02.10

(65)同一申请的已公布的文献号
申请公布号 CN 107430611 A

(43)申请公布日 2017.12.01

(30)优先权数据
62/114,691 2015.02.11 US

(85)PCT国际申请进入国家阶段日
2017.08.11

(86)PCT国际申请的申请数据
PCT/US2016/017263 2016.02.10

(87)PCT国际申请的公布数据
WO2016/130626 EN 2016.08.18

(73)专利权人 起元科技有限公司
地址 美国马萨诸塞州

(72)发明人 D·拉迪沃杰维奇

(74)专利代理机构 隆天知识产权代理有限公司 72003
代理人 张浴月 金鹏

(51)Int.Cl.
G06Q 10/06(2012.01)
G06F 16/26(2019.01)
G06F 16/25(2019.01)
G06F 16/901(2019.01)
G06Q 10/10(2012.01)
G06F 16/28(2019.01)

(56)对比文件
US 2010174692 A1,2010.07.08
CN 103678635 A,2014.03.26
CN 101321190 A,2008.12.10
毕春丽.信息通信技术标准化下的专利竞赛研究.《中国优秀博硕士学位论文全文数据库(博士)经济与管理科学辑》.2007,(第06期),J145-10.
Christoph Böhm 等.GovWILD:Integrating Open Government Data for Transparency.《WWW ″12 Companion: Proceedings of the 21st International Conference on World Wide Web》.2012,321-324.

审查员 李梦颖

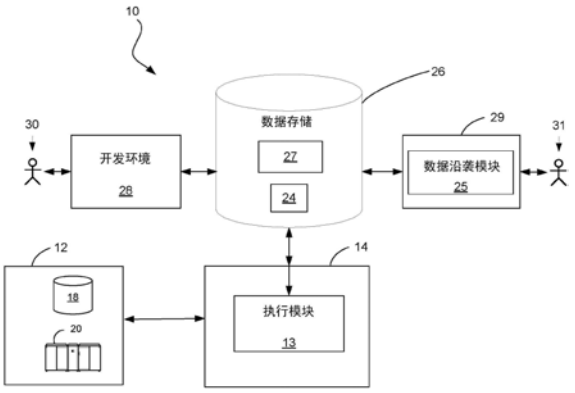
权利要求书4页 说明书15页 附图18页

(54)发明名称

过滤数据沿袭图

(57)摘要

管理沿袭信息包括处理有向图以将多个节点(102、104、106)与用于处理数据沿袭表示请求的信息相关联。处理包括:识别所述有向图中与数据存储器中存储的标准化数据元素对应的一个或多个节点(1362、1364、1366)的第一集合,并将从数据存储器中取回的数据元素去标准化;以及将连接至一个或多个节点的第一集合的第一组多个节点(1370、1372、1374)以及连接至一个或多个节点的第一集合的第二组多个节点(1376、1378、1380)与具有多个可能标签值的至少一个标签标识符相关联,其中可能标签值的数量至少与被标准化后的数据元素的数量一样多,其中在去标准化后的记录中表示不同数据元素的节点与所述标签标识符的不同值关联。



CN 107430611 B

1.一种用于在计算系统中管理沿袭信息的方法,所述方法包括:

在数据存储系统中储存有向图的规范,所述有向图的规范包括表示计算的多个节点以及表示在计算执行期间通过计算接收或产生的数据元素的多个节点以及表示计算和数据元素之间的沿袭关系的节点之间的有向链路;

使用至少一个处理器处理所述规范,以将多个节点与用于处理数据沿袭的表示的请求的信息相关联,所述处理包括:

识别所述有向图中与数据存储器中存储的标准化数据元素对应的一个或多个节点的第一集合,并将从所述数据存储器中取回的数据元素去标准化,其中将数据元素标准化包括将对应于多个数据元素的记录转换成具有针对至少一个字段的通用格式的多个记录,其中将数据元素去标准化包括将具有针对至少一个字段的通用格式的多个记录转换成对应于多个数据元素的单个记录;以及

将通过表示第一沿袭关系的第一有向链路连接至一个或多个节点的所述第一集合的第一组多个节点以及通过表示第二沿袭关系的第二有向链路连接至一个或多个节点的所述第一集合的第二组多个节点与具有多个可能标签值的至少一个标签标识符相关联,其中所述可能标签值的数量至少与被标准化后的数据元素的数量一样多,其中在去标准化后的记录中表示不同数据元素的节点与所述标签标识符的不同值关联。

2.根据权利要求1所述的方法,其中,由一个或多个节点的所述第一集合中的一个或多个节点执行的计算包括:

接收第一记录和第二记录,所述第一记录包括与所述第一组多个节点中的第一节点相关联的第一字段,所述第一字段存储有第一类型的值,所述第二记录包括与所述第一组多个节点中的第二节点相关联的第二字段,所述第二字段存储有所述第一类型的值;

生成标准化后的记录,所述标准化后的记录包括第三字段以及第四字段,所述第三字段存储有来自所述第一字段的值和来自所述第二字段的值,所述第四字段存储有用于标识是所述第一字段还是所述第二字段的值;以及

将所述标准化后的记录存储在所述数据存储器中。

3.根据权利要求2所述的方法,其中,由所述一个或多个节点的所述第一集合中的一个或多个节点执行的计算包括:

从所述数据存储器取回所述标准化后的记录;以及

至少部分基于所述第四字段中的值生成第一去标准化后的记录和第二去标准化后的记录,所述第一去标准化后的记录包括与所述第二组多个节点中的第三节点相关联的第五字段,所述第五字段存储来自所述第三字段的所述第一类型的值,所述第二去标准化后的记录包括与所述第二组多个节点中的第四节点相关联的第六字段,所述第六字段存储来自所述第三字段的所述第一类型的值。

4.根据权利要求3所述的方法,其中,一个或多个节点的所述第一集合包括:用于表示被执行以生成所述标准化后的记录的计算部分的节点;用于表示包括所述标准化后的记录的存储在所述数据存储器中的数据元素的节点;以及用于表示被执行以生成所述第一去标准化后的记录和所述第二去标准化后的记录的计算部分的节点。

5.根据权利要求2所述的方法,其中:

将所述第一组多个节点与所述标签标识符的不同值相关联包括将所述第一节点与所

述标签标识符的第一值相关联，以及将所述第二节点与所述标签标识符的第二值相关联，所述标签标识符的第二值与所述标签标识符的第一值不同；以及

将所述第二组多个节点与所述标签标识符的不同值相关联包括将第三节点与所述标签标识符的第一值相关联，以及将第四节点与所述标签标识符的第二值相关联。

6.根据权利要求5所述的方法，还包括：使用至少一个处理器处理针对所述第三节点的数据沿袭的表示的请求，所述处理包括：

确定所述第三节点与所述至少一个标签标识符之间的关联关系，确定所述标签标识符的所述多个可能标签值中的至少一个且少于全部的所述多个可能标签值的第一子集，所述第一子集包括所述标签标识符的所述第一值；

沿着从所述第三节点开始的有向链路的第一沿袭路径遍历多个节点，以确定所述第三节点的数据沿袭，其中确定所述数据沿袭包括：对于每个被遍历的节点，至少部分基于与所述被遍历的节点相关联的任何标签标识符或者标签值，确定是将所述被遍历的节点添加至所述数据沿袭还是将所述被遍历的节点从所述数据沿袭排除。

7.根据权利要求6所述的方法，其中，确定所述数据沿袭包括：

遍历所述第一沿袭路径上的所述第一节点；

响应于确定所述第一节点与所述标签标识符相关联并且与包括在所述第一子集中的所述标签标识符的所述第一值相关联，将所述第一节点添加到所述数据沿袭中；

沿着从所述第三节点开始的有向链路的第二沿袭路径遍历节点，以确定所述第三节点的数据沿袭，包括遍历所述第二沿袭路径上的所述第二节点；以及

响应于确定所述第二节点与所述标签标识符相关联但不与包括在所述第一子集中的任何标签标识符相关联，将所述第二节点从所述数据沿袭中排除。

8.根据权利要求7所述的方法，其中，确定所述数据沿袭还包括遍历一个或多个节点的所述第一集合，以及响应于确定一个或多个节点的所述第一集合不与所述标签标识符相关联而将一个或多个节点的所述第一集合添加到所述数据沿袭中。

9.根据权利要求6所述的方法，其中，确定所述数据沿袭包括：

对于每个被遍历的节点：

如果所述被遍历的节点不与所述标签标识符相关联，或者所述被遍历的节点与所述标签标识符相关联并且与所述第一子集中的至少一个标签值相关联，则将所述被遍历的节点添加到所述数据沿袭中；或

如果所述被遍历的节点与所述标签标识符相关联但不与所述标签标识符的所述多个可能标签值的第一子集中的至少一个标签值相关联，则将所述被遍历的节点从所述数据沿袭中排除。

10.根据权利要求9所述的方法，还包括：在将被遍历的节点从所述数据沿袭中排除时，终止所述第一沿袭路径的遍历。

11.根据权利要求6所述的方法，还包括：在遍历所述多个节点中的节点之前，分析所述多个节点，以确定与所述标签标识符相关联但不与所述标签标识符的所述多个可能标签值的所述第一子集中的至少一个标签值相关联的所述多个节点的第一子集。

12.根据权利要求11所述的方法，其中，确定是将所述被遍历的节点添加到所述数据沿袭还是将所述被遍历的节点从所述数据沿袭中排除包括：确定所述被遍历的节点是否包括

在所述多个节点的所述第一子集中。

13.根据权利要求12所述的方法，其中，确定是否将所述被遍历的节点添加到所述数据沿袭包括：如果所述被遍历的节点没有包括在所述多个节点的所述第一子集中，则将所述被遍历的节点添加到所述数据沿袭，如果所述被遍历的节点包括在所述多个节点的所述第一子集中，则将所述被遍历的节点从所述数据沿袭中排除。

14.根据权利要求1所述的方法，其中所述多个节点中的至少一些节点嵌套在所述多个节点中的其他节点内。

15.根据权利要求14所述的方法，其中，相对于与节点嵌套层级中较高级别的节点相关联的标签标识符及其相关联的标签值的子集，与所述节点嵌套层级中较低级别的节点相关联的标签标识符及其相关联的标签值的子集具有优先级。

16.根据权利要求6所述的方法，其中，沿着所述第一沿袭路径遍历节点包括：从所述第三节点开始沿着与所述有向链路的方向相反的方向遍历节点。

17.根据权利要求1所述的方法，还包括：生成所述有向图的规范，其中所述多个节点中的一个或多个节点表示计算，所述多个节点中的一个或多个节点表示在计算执行期间通过计算接收或产生的数据元素，所述有向链路表示计算和数据元素之间的沿袭关系。

18.根据权利要求17所述的方法，其中，所述多个节点中的一个或多个节点表示所述多个节点中的一个或多个其他节点的分组。

19.根据权利要求6所述的方法，其中，所述标签标识符的所述多个可能标签值的所述第一子集由与所述第三节点相关联的一个或多个标签值构成。

20.根据权利要求6所述的方法，其中，所述标签标识符的所述多个可能标签值的所述第一子集由从用户输入接收到的一个或多个标签值构成。

21.根据权利要求6所述的方法，还包括：沿着从所述第三节点开始的有向链路的多个沿袭路径遍历节点，以确定所述第三节点的数据沿袭，其中确定数据沿袭包括：对于每个被遍历的节点，至少部分基于与所述被遍历的节点相关联的任何标签标识符或标签值，确定是将所述被遍历的节点添加到所述数据沿袭还是将所述被遍历的节点从所述数据沿袭中排除。

22.根据权利要求6所述的方法，还包括：默认将所述第三节点添加到所述数据沿袭。

23.一种储存在计算机可读介质上的软件，用于在计算系统中管理沿袭信息，所述软件包括用于使计算系统执行以下操作的指令：

在数据存储系统中储存有向图的规范，所述有向图的规范包括表示计算的多个节点以及表示在计算执行期间通过计算接收或产生的数据元素的多个节点以及表示计算和数据元素之间的沿袭关系的节点之间的有向链路；

使用至少一个处理器处理所述规范，以将多个节点与用于处理数据沿袭表示请求的信息相关联，所述处理包括：

识别所述有向图中与数据存储器中存储的标准化数据元素对应的一个或多个节点的第一集合，并将从所述数据存储器中取回的数据元素去标准化，其中将数据元素标准化包括将对应于多个数据元素的记录转换成具有针对至少一个字段的通用格式的多个记录，其中将数据元素去标准化包括将具有针对至少一个字段的通用格式的多个记录转换成对应于多个数据元素的单个记录；以及

将通过表示第一沿袭关系的第一有向链路连接至一个或多个节点的所述第一集合的第一组多个节点以及通过表示第二沿袭关系的第二有向链路连接至一个或多个节点的所述第一集合的第二组多个节点与具有多个可能标签值的至少一个标签标识符相关联，其中所述可能标签值的数量至少与被标准化后的数据元素的数量一样多，其中在去标准化后的记录中表示不同数据元素的节点与所述标签标识符的不同值关联。

24.一种计算系统，用于管理沿袭信息，所述计算系统包括：

数据存储系统，用于储存有向图的规范，所述有向图的规范包括表示计算的多个节点以及表示在计算执行期间通过计算接收或产生的数据元素的多个节点以及表示计算和数据元素之间的沿袭关系的节点之间的有向链路；

至少一个处理器，配置为处理所述规范，以将多个节点与用于处理数据沿袭表示请求的信息相关联，所述处理包括：

识别所述有向图中与数据存储器中存储的标准化数据元素对应的一个或多个节点的第一集合，并将从所述数据存储器中取回的数据元素去标准化，其中将数据元素标准化包括将对应于多个数据元素的记录转换成具有针对至少一个字段的通用格式的多个记录，其中将数据元素去标准化包括将具有针对至少一个字段的通用格式的多个记录转换成对应于多个数据元素的单个记录；以及

将通过表示第一沿袭关系的第一有向链路连接至一个或多个节点的所述第一集合的第一组多个节点以及通过表示第二沿袭关系的第二有向链路连接至一个或多个节点的所述第一集合的第二组多个节点与具有多个可能标签值的至少一个标签标识符相关联，其中所述可能标签值的数量至少与被标准化后的数据元素的数量一样多，其中在去标准化后的记录中表示不同数据元素的节点与所述标签标识符的不同值关联。

过滤数据沿袭图

[0001] 相关申请的交叉引用

[0002] 本申请要求于2015年2月11日提交的序号为62/114,691的美国申请的优先权。

技术领域

[0003] 本说明书涉及用于过滤数据沿袭(data lineage)图的沿袭标签。

背景技术

[0004] 在数据处理系统中,某些类型的用户通常期望能够访问贯穿系统的数据沿袭的视觉表示。这样的“数据沿袭图”可以包括系统中的数据和用于处理该数据实体的图形表示以及它们之间的依赖关系。通常,在许多用途中,可以使用这些数据沿袭图来降低风险,验证合规性义务(regulatory compliance obligation),提高业务流程效率以及保护数据。重要的是数据沿袭图是正确且完整的。

[0005] 能够产生并显示数据沿袭图的一些系统能够自动呈现端到端的数据沿袭图,这些数据沿袭图示出数据项的表示,并且显示表示了用于消耗或生成这些数据项的处理实体的项的表示。在一些示例中,特定项上游的沿袭路径有时称为该项的“依赖性沿袭(dependency lineage)”,并且特定项下游的沿袭路径有时称为该项的“影响沿袭(impact lineage)”。在其他示例中,特定项上游的沿袭路径被称为“沿袭”或“依赖性”,并且特定项下游的沿袭路径称为“影响”或“影响分析”。如本文所使用的,“数据沿袭图”可以包括相对于任意给定项的上游依赖性沿袭和/或下游影响性沿袭。能够生成并显示数据沿袭图的一些系统允许用户将数据沿袭图中项的逻辑和/或物理组分解成单个元素。能够生成并显示数据沿袭图的一些系统能够通过丰富的数据信息(如数据质量评分)来增强数据沿袭图。

发明内容

[0006] 在一个方面,通常,一种用于在计算系统中管理沿袭信息的方法,所述方法包括:在数据存储系统中储存有向图的规范,所述有向图的规范包括表示计算的多个节点以及表示在计算执行期间通过计算接收或产生的数据元素的多个节点以及表示计算和数据元素之间的沿袭关系的节点之间的有向链路;使用至少一个处理器处理所述规范,以将多个节点与用于处理数据沿袭表示请求的信息相关联。所述处理包括:识别所述有向图中与数据存储器中存储的标准化数据元素对应的一个或多个节点的第一集合,并将从所述数据存储器中取回的数据元素去标准化,其中将数据元素标准化包括将对应于多个数据元素的记录转换成具有针对至少一个字段的通用格式的多个记录,其中将数据元素去标准化包括将具有针对至少一个字段的通用格式的多个记录转换成对应于多个数据元素的单个记录;以及将通过表示第一沿袭关系的第一有向链路连接至一个或多个节点的所述第一集合的第一组多个节点以及通过表示第二沿袭关系的第二有向链路连接至一个或多个节点的所述第一集合的第二组多个节点与具有多个可能标签值的至少一个标签标识符相关联,其中所述可能标签值的数量至少与被标准化后的数据元素的数量一样多,其中在去标准化后的记录

中表示不同数据元素的节点与所述标签标识符的不同值关联。

[0007] 多个方面可以包括如下特征中的一个或多个。

[0008] 一个或多个节点的所述第一集合中的一个或多个节点执行的计算包括:接收第一记录和第二记录,所述第一记录包括与所述第一组多个节点中的第一节点相关联的第一字段,所述第一字段存储有第一类型的值,所述第二记录包括与所述第一组多个节点中的第二节点相关联的第二字段,所述第二字段存储有所述第一类型的值;生成标准化后的记录,所述标准化后的记录包括第三字段以及第四字段,所述第三字段存储有来自所述第一字段的值和来自所述第二字段的值,所述第四字段存储有用于标识是所述第一字段还是所述第二字段的值;以及将所述标准化后的记录存储在所述数据存储器中。

[0009] 所述一个或多个节点的所述第一集合中的一个或多个节点执行的计算包括:从所述数据存储器取回所述标准化后的记录;以及至少部分基于所述第四字段中的值生成第一去标准化后的记录和第二去标准化后的记录,所述第一去标准化后的记录包括与所述第二组多个节点中的第三节点相关联的第五字段,所述第五字段存储来自所述第三字段的所述第一类型的值,所述第二去标准化后的记录包括与所述第二组多个节点中的第四节点相关联的第六字段,所述第六字段存储来自所述第三字段的所述第一类型的值。

[0010] 一个或多个节点的所述第一集合包括:用于表示被执行以生成所述标准化后的记录的计算部分的节点;用于表示包括所述标准化后的记录的存储在所述数据存储器中的数据元素的节点;以及用于表示被执行以生成所述第一去标准化后的记录和所述第二去标准化后的记录的计算部分的节点。

[0011] 将所述第一组多个节点与所述标签标识符的不同值相关联包括将所述第一节点与所述标签标识符的第一值相关联,以及将所述第二节点与所述标签标识符的第二值相关联,所述标签标识符的第二值与所述标签标识符的第一值不同;以及将所述第二组多个节点与所述标签标识符的不同值相关联包括将所述第三节点与所述标签标识符的第一值相关联,以及将所述第四节点与所述标签标识符的第二值相关联。

[0012] 所述方法还包括:使用至少一个处理器处理针对所述第三节点的数据沿袭表示请求,所述处理包括:确定所述第三节点与所述至少一个标签标识符之间的关联关系,确定第一子集,所述第一子集是所述标签标识符的所述多个可能标签值中的至少一个且少于全部所述多个可能标签值;沿着从所述第三节点开始的有向链路的第一沿袭路径遍历多个节点,以确定所述第三节点的数据沿袭,其中确定所述数据沿袭包括:对于每个被遍历的节点,至少部分基于与所述被遍历的节点相关联的任何标签标识符或者标签值,确定是将所述被遍历的节点添加至所述数据沿袭还是将所述被遍历的节点从所述数据沿袭排除。

[0013] 确定所述数据沿袭包括:遍历所述第一沿袭路径上的所述第一节点;响应于确定所述第一节点与所述标签标识符相关联并且与包括在所述第一子集中的所述标签标识符的所述第一值相关联,将所述第一节点添加到所述数据沿袭中;沿着从所述第三节点开始的有向链路的第二沿袭路径遍历节点,以确定所述第三节点的数据沿袭,包括遍历所述第二沿袭路径上的所述第二节点;以及响应于确定所述第二节点与所述标签标识符相关联但不与包括在所述第一子集中的任何标签标识符相关联,将所述第二节点从所述数据沿袭中排除。

[0014] 确定所述数据沿袭还包括遍历一个或多个节点的所述第一集合,以及响应于确定

一个或多个节点的所述第一集合不与所述标签标识符相关联而将一个或多个节点的所述第一集合添加到所述数据沿袭中。

[0015] 确定所述数据沿袭包括:对于每个被遍历的节点:如果所述被遍历的节点不与所述标签标识符相关联,或者所述被遍历的节点与所述标签标识符相关联并且与所述第一子集中的至少一个标签值相关联,则将所述被遍历的节点添加到所述数据沿袭中;或如果所述被遍历的节点与所述标签标识符相关联但不与所述标签标识符的所述多个可能标签值的第一子集中的至少一个标签值相关联,则将所述被遍历的节点从所述数据沿袭中排除。

[0016] 所述方法还包括:在将被遍历的节点从所述数据沿袭中排除时,终止所述第一沿袭路径的遍历。

[0017] 所述方法还包括:在遍历所述多个节点中的节点之前,分析所述多个节点,以确定与所述标签标识符相关联但不与所述标签标识符的所述多个可能标签值的所述第一子集中的至少一个标签值相关联的所述多个节点的第一子集。

[0018] 确定是将所述被遍历的节点添加到所述数据沿袭还是将所述被遍历的节点从所述数据沿袭中排除包括:确定所述被遍历的节点是否包括在所述多个节点的所述第一子集中。

[0019] 确定是否将所述被遍历的节点添加到所述数据沿袭包括:如果所述被遍历的节点没有包括在所述多个节点的所述第一子集中,则将所述被遍历的节点添加到所述数据沿袭,如果所述被遍历的节点包括在所述多个节点的所述第一子集中,则将所述被遍历的节点从所述数据沿袭中排除。

[0020] 所述多个节点中的至少一些节点嵌套在所述多个节点中的其他节点内。

[0021] 相对于与节点嵌套层级中较高级别的节点相关联的标签标识符及其相关联的标签值的子集,与所述节点嵌套层级中较低级别的节点相关联的标签标识符及其相关联的标签值的子集具有优先级。

[0022] 沿着所述第一沿袭路径遍历节点包括:从所述第三节点开始沿着与所述有向链路的方向相反的方向遍历节点。

[0023] 所述方法还包括:生成所述有向图的规范,其中所述多个节点中的一个或多个节点表示计算,所述多个节点中的一个或多个节点表示在计算执行期间通过计算接收或产生的数据元素,所述有向链路表示计算和数据元素之间的沿袭关系。

[0024] 所述多个节点中的一个或多个表示所述多个节点中的一个或多个其他节点的分组。

[0025] 所述标签标识符的所述多个可能标签值的所述第一子集由与所述第三节点相关联的一个或多个标签值构成。

[0026] 所述标签标识符的所述多个可能标签值的所述第一子集由从用户输入接收到的一个或多个标签值构成。

[0027] 所述方法还包括:沿着从所述第三节点开始的有向链路的多个沿袭路径遍历节点,以确定所述第三节点的数据沿袭,其中确定数据沿袭包括:对于每个被遍历的节点,至少部分基于与所述被遍历的节点相关联的任何标签标识符或标签值,确定是将所述被遍历的节点添加到所述数据沿袭还是将所述被遍历的节点从所述数据沿袭中排除。

[0028] 所述方法还包括:默认将所述第三节点添加到所述数据沿袭。

[0029] 在另一方面中，一种储存在计算机可读介质上的软件，用于在计算系统中管理沿袭信息。所述软件包括用于使计算系统执行以下操作的指令：在数据存储系统中储存有向图的规范，所述有向图的规范包括表示计算的多个节点以及表示在计算执行期间通过计算接收或产生的数据元素的多个节点以及表示计算和数据元素之间的沿袭关系的节点之间的有向链路；使用至少一个处理器处理所述规范，以将多个节点与用于处理数据沿袭表示请求的信息相关联，所述处理包括：识别所述有向图中与数据存储器中存储的标准化数据元素对应的一个或多个节点的第一集合，并将从所述数据存储器中取回的数据元素去标准化，其中将数据元素标准化包括将对应于多个数据元素的记录转换成具有针对至少一个字段的通用格式的多个记录，其中将数据元素去标准化包括将具有针对至少一个字段的通用格式的多个记录转换成对应于多个数据元素的单个记录；以及将通过表示第一沿袭关系的第一有向链路连接至一个或多个节点的所述第一集合的第一组多个节点以及通过表示第二沿袭关系的第二有向链路连接至一个或多个节点的所述第一集合的第二组多个节点与具有多个可能标签值的至少一个标签标识符相关联，其中所述可能标签值的数量至少与被标准化后的数据元素的数量一样多，其中在去标准化后的记录中表示不同数据元素的节点与所述标签标识符的不同值关联。

[0030] 在另一方面中，一种计算系统，用于管理沿袭信息，所述计算系统包括：数据存储系统，用于储存有向图的规范，所述有向图的规范包括表示计算的多个节点以及表示在计算执行期间通过计算接收或产生的数据元素的多个节点以及表示计算和数据元素之间的沿袭关系的节点之间的有向链路；至少一个处理器，配置为处理所述规范，以将多个节点与用于处理数据沿袭表示请求的信息相关联。所述处理包括：识别所述有向图中与数据存储器中存储的标准化数据元素对应的一个或多个节点的第一集合，并将从所述数据存储器中取回的数据元素去标准化，其中将数据元素标准化包括将对应于多个数据元素的记录转换成具有针对至少一个字段的通用格式的多个记录，其中将数据元素去标准化包括将具有针对至少一个字段的通用格式的多个记录转换成对应于多个数据元素的单个记录；以及将通过表示第一沿袭关系的第一有向链路连接至一个或多个节点的所述第一集合的第一组多个节点以及通过表示第二沿袭关系的第二有向链路连接至一个或多个节点的所述第一集合的第二组多个节点与具有多个可能标签值的至少一个标签标识符相关联，其中所述可能标签值的数量至少与被标准化后的数据元素的数量一样多，其中在去标准化后的记录中表示不同数据元素的节点与所述标签标识符的不同值关联。

[0031] 在另一方面中，通常，一种计算系统，用于管理沿袭信息，所述计算系统包括：用于存储有向图的规范的装置，所述有向图的规范包括表示计算的多个节点以及表示在计算执行期间通过计算接收或产生的数据元素的多个节点以及表示计算和数据元素之间的沿袭关系的节点之间的有向链路；以及用于处理所述规范以将多个节点与用于处理数据沿袭表示请求的信息相关联的装置。所述处理包括：识别所述有向图中与数据存储器中存储的标准化数据元素对应的一个或多个节点的第一集合，并将从所述数据存储器中取回的数据元素去标准化，其中将数据元素标准化包括将对应于多个数据元素的记录转换成具有针对至少一个字段的通用格式的多个记录，其中将数据元素去标准化包括将具有针对至少一个字段的通用格式的多个记录转换成对应于多个数据元素的单个记录；以及将通过表示第一沿袭关系的第一有向链路连接至一个或多个节点的所述第一集合的第一组多个节点以及通

过表示第二沿袭关系的第二有向链路连接至一个或多个节点的所述第一集合的第二组多个节点与具有多个可能标签值的至少一个标签标识符相关联，其中所述可能标签值的数量至少与被标准化后的数据元素的数量一样多，其中在去标准化后的记录中表示不同数据元素的节点与所述标签标识符的不同值关联。

[0032] 多个方面可以包括一个或多个下述优点。

[0033] 在一些常规方案中，本申请所述方法适用于用户已经用沿袭标签形式的附加信息来增强了各节点的沿袭图。沿袭标签用于过滤沿袭图，使得可以定位特定的用户定义维度，而不会将沿袭仅仅限制为与通过过滤所挑选的对象相邻的对象，或者通过过滤所挑选的对象的直系沿袭或影响路径内的对象。

[0034] 相比于其他技术(例如，简单过滤或沿袭汇总)，本申请所述的方法可以以更具选择性的方式降低复杂度。例如，一些沿袭图中的一个复杂度的来源是执行标准化和去标准化操作的节点，使用沿袭标签可以有助于降低这类沿袭图中的复杂度，如本申请更详细描述的那样。

[0035] 这些方法通过允许用户定义不同类型的标签来实现选择性，每个标签都有自己的一组枚举值。每种类型的标签充当其自己的独立过滤维度。标签也可以基于沿袭图的配置而自动定义，例如执行标准化和去标准化的节点。

[0036] 在一些示例中，用于选择要将哪些节点包括在过滤后的数据沿袭图中以及从过滤后的数据沿袭图中排除哪些节点的算法遍及沿袭图的各节点，并通过将所遇到的标签(与标签相关联的任何节点)与一组活动标签(active tag)进行比较来确定要排除哪个或哪个。

[0037] 除了其他优点之外，本申请描述的方法并不是简单地将图限制成具有活动标签的项的上游和下游节点。相反，未标注标签的节点也保留在图中。因此，这些方法允许手术去除枚举组的非期望元素的上游或下游的节点。这种选择性有利地提供了解决通过简单过滤或者沿袭汇总(例如，数据的标准化/去标准化或超载对象(overloaded object))而无法解决的问题的能力。

[0038] 随着数据处理系统复杂度的增加，数据沿袭图也变得越来越复杂，在单一视图中呈现出由相应图标所表示的许多节点(例如，数据节点和数据变换节点，统称“节点”)。随着在数据沿袭图中呈现给用户的节点数量的增加，数据沿袭图也可能变得难以理解。

[0039] 在一些示例中，数据处理图包括许多部分重叠的数据沿袭路径，它们中只有一小部分是用户通常所感兴趣的。例如，在数据处理图的上游部分中，存在许多不同的系统，这些系统基于例如地理区域来处理数据。在数据处理图的中游部分中，对来自该图的上游部分的数据进行分组，并且在该图的下游部分中，处理已分组的数据。用户可能有兴趣分析单一地理区域的沿袭路径。

[0040] 本文所述的方法利用了一种认识，即在给定的数据处理系统内，相互关联的某些节点通常就是用户查看时所感兴趣的节点。在本文所述的方法中，使用沿袭标签来关联彼此相关的节点集。然后可以基于沿袭标签来过滤节点，以允许用户生成过滤后的数据沿袭图，其中仅包括用户查看时所感兴趣的节点。

[0041] 基于以下描述和权利要求，本发明的其他特征和优点将变得明显。

附图说明

[0042] 图1是包括沿袭过滤特征的数据处理系统的框图。

[0043] 图2a是包括了带有沿袭标签的节点的数据沿袭图。

[0044] 图2b是“区域:US”的第一过滤后的数据沿袭图。

[0045] 图3是过滤算法的框图。

[0046] 图4是图2a的数据沿袭图,其中选择了一个节点并将几个节点添加至黑名单。

[0047] 图5是图2a的数据沿袭图,其中遍历了第一数据沿袭路径。

[0048] 图6是图2a的数据沿袭图,其中遍历了第二数据沿袭路径。

[0049] 图7是图2a的数据沿袭图,其中遍历了第三数据沿袭路径。

[0050] 图8是图2a的数据沿袭图,其中遍历了第四数据沿袭路径。

[0051] 图9是图2a的数据沿袭图,其中遍历了第五数据沿袭路径。

[0052] 图10是图2a的数据沿袭图,其中遍历了第六数据沿袭路径。

[0053] 图11是图2a的数据沿袭图,其中排除了无关联的节点。

[0054] 图12是“区域:US、MEX”的第二过滤后的数据沿袭图。

[0055] 图13是计算机程序的数据沿袭图,该计算机程序被配置为执行数据标准化和去标准化。

[0056] 图14是图13的数据沿袭图,其包括与一些节点相关联的沿袭标签。

[0057] 图15是图13的数据沿袭图,其中某些节点被添加至黑名单。

[0058] 图16是基于沿袭标签的过滤过程中的图13的数据沿袭图。

[0059] 图17是过滤后的数据沿袭图。

具体实施方式

[0060] 图1示出了数据处理系统10的示例,其可以使用基于沿袭标签的数据沿袭过滤技术。系统10包括数据源12,该数据源12可以包括一个或多个数据源,例如存储设备或者至在线数据流的连接,每个数据源可以存储或提供各种格式中任意格式的数据(例如,数据库表、电子表格文件、纯文本文件或大型机使用的本地格式)。执行环境14包括执行模块13。执行环境14可以被托管在一个或多个通用计算机上,该一个或多个通用计算机受合适的操作系统的控制,例如某版本的UNIX操作系统。例如,执行环境14可包括多节点并行计算环境,该多节点并行计算环境包括使用多个中央处理单元(CPU)或处理器内核的计算机系统的配置,其为本地的(例如,诸如对称多处理(SMP)计算机之类的多处理器系统),或为本地分布式(例如,耦接为集群或大规模并行处理(MPP)系统的多个处理器,或为远程的,或为远程分布式的(例如,经由局域网(LAN)和/或广域网(WAN)耦接的多个处理器),或为其任意组合。

[0061] 提供了数据源12的存储设备可以对于执行环境14来说是本地的,例如,存储在连接到托管有执行环境14的计算机的存储介质(例如,硬盘驱动器18)上,或者可以对于执行环境14来说是远程的,例如,通过远程连接(例如,由云计算基础设施提供的远程连接)被托管在与托管了执行环境14的计算机进行通信的远程系统(例如,大型机20)上)。

[0062] 系统10包括开发环境28,在该开发环境28中,开发者30能够指定数据处理计算机程序27(例如,数据流图表),并将程序存储在执行环境14可访问的数据存储系统26中。执行模块13根据计算机程序27来处理来自数据源的数据,从而产生输出数据24。输出数据24可

以被存储回数据源12中或被存储在数据存储系统26中，或者被使用。在一些实施方式中，开发环境28是用于将应用程序开发为数据流图表的系统，该数据流图表包括通过多个顶点之间的有向链路(表示工作元素的流，即数据)连接的顶点(表示数据处理组件或数据集)。例如，这种环境在美国公开号为2007/0011668的标题为“管理基于图形的应用的参数”的美国专利公开文件中有更详细的描述，该专利通过引用的方式并入本申请。在标题为“EXECUTING COMPUTATIONS EXPRESSED AS GRAPHS(执行表达为图的计算)”的美国专利5,966,072中描述了一种用于执行这种基于图的计算的系统，其通过引用的方式并入本申请。根据该系统制作的数据流图提供了用于将信息输入和输出由图形组件表示的各个进程、用于在进程之间移动信息以及用于定义进程的运行顺序的方法。该系统包括从任何可用方法选择进程间通信方法的算法(例如，根据图的链路的通信沿袭路径可以使用TCP/IP或UNIX域套接字，或者使用共享存储器在进程之间传递数据)。

[0063] 系统10包括企业环境29，用户31(例如，企业用户或数据架构师)可以通过该企业环境29请求和查看数据沿袭图。为了生成数据沿袭图，企业环境29包括分析系统元数据以生成数据沿袭图的数据沿袭模块25，系统元数据包括表征表示对应于计算机程序27不同部分的计算的数据变换节点的元数据(例如，不同的数据流图或数据流图内的不同组件)以及表征由计算机程序27访问或生成的数据节点的元数据(例如，来自数据源12的数据集或对应于输出数据24的数据集)。

[0064] 在一些示例中，表征数据变换节点和/或数据节点的元数据包括数据沿袭标签(有时称为“沿袭标签”)。通常，每个沿袭标签都具有标签标识符和许多可能的标签值。例如，一个示例性沿袭标签可以具有的标签标识符为“区域”以及可能的标签值为“美国、英国、墨西哥、加拿大”。

[0065] 计算机程序27中的至少一些节点(即，数据变换节点和/或数据节点)与一个或多个沿袭标签和该沿袭标签的可能标签值的子集相关联，该子集包括所述一个或多个沿袭标签的可能标签值的至少一个值并且少于所述一个或多个沿袭标签的所有可能的标签值。

[0066] 参考图2a，计算机程序27的示例性端到端数据沿袭图100包括多个数据节点102、多个数据变换节点104和多个容器节点106，它们通过有向链路108互连。

[0067] 特别地，图100包括第一数据变换节点110、第二数据变换节点112、第三数据变换节点114、第四数据变换节点116、第五数据变换节点118、第六数据变换节点120、第七数据变换节点122、第八数据变换节点124、第九数据变换节点126、第十数据变换节点128、第十一数据变换节点130、第十二数据变换节点132和第十三数据变换节点134。图100包括第一数据节点136和第二数据节点138。图100包括第一容器节点140、第二容器节点142、第三容器节点144和第四容器节点146。

[0068] 第一数据变换节点110与“区域”沿袭标签相关联。在该示例中，“区域”沿袭标签具有三个可能的标签值：“US、CA、MEX”。第一变换节点110与包括“US”的“区域”沿袭标签的可能标签值的子集相关联。第三数据变换节点114与“区域”沿袭标签相关联，并且与包括“MEX”的“区域”沿袭标签的可能标签值的子集相关联。第四数据变换节点116与“区域”沿袭标签相关联标签，并且与包括“US”的“区域”沿袭标签的可能标签值的子集相关联。第五数据变换节点118与“区域”沿袭标签相关联，并且与包括“CA”的“区域”沿袭标签的可能标签值的子集相关联。第六数据变换节点120与“区域”沿袭标签相关联，并且与包括“US”的“区

域”沿袭标签的可能标签值的子集相关联。

[0069] 第七数据变换节点122与“区域”沿袭标签相关联，并且与包括“CA”的“区域”沿袭标签的可能标签值的子集相关联。第八数据变换节点124与“区域”沿袭标签相关联，并且与包括“MEX”的“区域”沿袭标签的可能标签值的子集相关联。第九数据变换节点126与“区域”沿袭标签相关联，并且与包括“US”的“区域”沿袭标签的可能标签值的子集相关联。第十三数据变换节点134与“区域”沿袭标签相关联，并且与包括“CA”的“区域”沿袭标签的可能标签值的子集相关联。

[0070] 第一数据节点136与“区域”沿袭标签相关联，并且与包括“US,MEX”的“区域”沿袭标签的可能标签值的子集相关联。第二数据节点138与“区域”沿袭标签相关联，并且与包括“CA”的“区域”沿袭标签的可能标签值的子集相关联。

[0071] 1过滤后的数据沿袭生成

[0072] 在一些示例中，图2a的端到端数据沿袭图100由数据沿袭模块25处理，以产生过滤后的数据沿袭图。参考图2b，过滤后的数据沿袭图201的一个示例是根据“区域”沿袭标签和“US”沿袭标签值而进行过滤的端到端数据沿袭图100的一个版本。从图中可以看出，与图2a的端到端数据沿袭图100相比，过滤后的数据沿袭图201大大简化，因为从数据沿袭中排除了与“区域”沿袭标签标识符相关联的并且具有仅包括“CA”或仅包括“MEX”的沿袭标签值的多个节点。

[0073] 如下文参考图3至图12所详细描述的那样，为了实现诸如图12的第一过滤后数据沿袭图200之类的过滤后数据沿袭图，用户选择节点并命令数据沿袭模块25确定该节点的过滤后数据沿袭。数据沿袭模块25首先确定所选节点与一个或多个标签标识符之间的关联。对于与所选节点相关联的每个标签标识符，数据沿袭模块25确定标签标识符的可能标签值中的哪些标签值与所选节点相关联。通常，用户可以选择任意数量的要与所选节点相关联的标签值。所选标签值可以表示可能的标签值的子集，例如单个标签值(例如，如图2b示例所示的“US”)或多个标签值(例如，如图12示例所示的“US”和“MEX”)。通常，选择少于与标签标识符相关联的所有可能标签值的一些标签值。

[0074] 然后，数据沿袭模块25遍历从所选节点开始的一个或多个沿袭路径，以确定所选节点的过滤后数据沿袭。通常，每个沿袭路径包括由有向链路连接的一个或多个节点。当数据沿袭模块25遍历一个或多个节点时，基于与该一个或多个所遍历的节点相关联的任意标签标识符或标签值来决定是否将节点添加至过滤后的数据沿袭或者从过滤后的数据沿袭中排除节点。

[0075] 参考图3，在一些示例中，当数据沿袭模块25遍历沿袭路径时，过滤算法300用于确定沿着沿袭路径的节点是否应被标记为包括在过滤后的数据沿袭图中，或者是否应该被标记为从过滤后的数据沿袭图排除。

[0076] 特别地，当数据沿袭模块25到达沿袭路径上的被访问节点350时，向过滤算法300提供被访问节点350和所选节点352。

[0077] 在过滤算法300的第一步骤354中，执行测试，以确定与被访问节点350相关联的任意沿袭标签标识符是否与所选节点352相关联的沿袭标签标识符匹配。如果没有一个沿袭标签标识符匹配，则算法进行到第二步骤356，在该步骤中，被访问节点350被标记为包括在过滤后的数据沿袭图中。如果与被访问节点350和所选节点352两者相关联的一个或多个沿

袭标签标识符匹配，则算法进行到第三步骤358。

[0078] 在第三步骤358中，对于每个识别出来的匹配沿袭标签标识符，执行测试，以确定与被访问节点350相关联的针对该沿袭标签标识符的沿袭标签值中的一个或多个是否匹配与所选节点352相关联的针对该沿袭标签标识符的沿袭标签值中的一个或多个。如果与被访问节点350相关联的针对该沿袭标签标识符的沿袭标签值中的一个或多个匹配与所选节点352相关联的针对该沿袭标签标识符的沿袭标签值中的一个或多个，则算法进行到第二步骤356，在该步骤中，被访问节点350被标记为包括在过滤后的数据沿袭图中。

[0079] 或者，如果与被访问节点350相关联的针对该沿袭标签标识符的沿袭标签值中没有一个能够匹配与所选节点352相关联的针对该沿袭标签标识符的沿袭标签值中的任一个，则算法进行到第四步骤360，在该步骤中，被访问节点350被标记为排除在过滤后的数据沿袭图之外。

[0080] 在一些示例中，当数据沿袭模块25遍历了端到端数据沿袭中的沿袭路径时，执行过滤算法400。在其他示例中，一旦节点被选择用于数据沿袭计算，并且在遍历任意沿袭路径之前，将过滤算法应用于端到端数据沿袭中除所选节点之外的所有节点。过滤算法300标记为从过滤后的数据沿袭图中排除的任何节点被添加到“黑名单”。然后，数据沿袭模块25遍历从所选节点开始的沿袭路径。在由数据沿袭模块25所访问的每个节点处，数据沿袭模块25搜索黑名单以确定被访问节点是否存在于黑名单中。如果被访问节点存在于黑名单中，则终止沿袭路径的遍历，并且该被访问节点不包括在过滤后的数据沿袭中。

[0081] 参考图4，用户已经选择了图2a的端到端数据沿袭图100中的第一数据节点136以用于生成过滤后的数据沿袭。如上所述，第一数据节点136与“区域”沿袭标签标识符相关联，并且与包括“US、MEX”且不包括沿袭标签值“CA”的“区域”沿袭标签标识符的可能标签值的子集相关联。

[0082] 由于“区域”沿袭标签标识符的“CA”沿袭标签值不包括在与第一数据节点136相关联的沿袭标签值的子集中，所以数据沿袭模块25被配置为将既与“区域”沿袭标签标识符相关联并且还具有仅包括“CA”的标签值子集的节点从数据沿袭中排除。

[0083] 在一些示例中，在接收到第一数据节点136的选择之后，数据沿袭模块25处理端到端数据沿袭图100以生成黑名单。如上所述，检查除第一数据节点136之外的所有节点来识别与“区域”沿袭标签标识符相关联且具有仅包括“CA”的标签值子集的节点，从而生成黑名单。将过滤算法300应用到图4的端到端数据沿袭而导致将第五数据变换节点118、第七数据变换节点122、第十三数据变换节点134和第二数据节点138添加到黑名单。在图4中，用对角线图案填充被添加到黑名单的节点。

[0084] 一旦生成黑名单，则遍历从第一数据节点136开始的每条沿袭路径(即，由有向链路108连接的各节点)，直到访问到位于端到端沿袭图100的边缘的节点，或者访问到黑名单上的节点为止。在遇到位于端到端数据沿袭图100边缘的节点的情况下，在该节点之前的整条沿袭路径(包括该节点)被包括在过滤后的数据沿袭中。在遇到黑名单上的节点的情况下，终止沿袭路径的遍历，并且在黑名单上的该节点之前的沿袭路径(不包括该节点)被包括在过滤后的数据沿袭中。

[0085] 参考图5，数据沿袭模块25遍历第一数据节点136和第九数据变换节点126之间的第一沿袭路径。由于用户选择了第一数据节点136，所以它被默认标记为包括在过滤后的数

据沿袭中。接下来遍历第一数据节点136和第十数据变换节点128之间的有向链路108。然后搜索黑名单以确定黑名单是否包括第十数据变换节点128。由于第十数据变换节点128不在黑名单中，因此它被标记为包括在过滤后的数据沿袭中。接下来遍历第十数据变换节点128和第九数据变换节点126之间的有向链路108。然后搜索黑名单以确定黑名单是否包括第九数据变换节点126。由于第九数据变换节点126不在黑名单中，因此它被标记为包括在过滤后的数据沿袭中。由于第九数据变换节点126位于端到端沿袭图100的边缘，所以第一沿袭路径的遍历完成。

[0086] 参考图6，数据沿袭模块25遍历在第一数据节点136和第一数据变换节点110之间延伸的第二沿袭路径。由于先前已经遍历了第一数据节点136和第十数据变换节点128之间的有向链路108，所以数据沿袭模块25从遍历第十数据变换节点128和第四数据变换节点116之间的有向链路108开始其遍历。

[0087] 当沿袭模块25到达第四数据变换节点116时，搜索黑名单以确定黑名单是否包括第四数据变换节点116。由于第四数据变换节点116不在黑名单中，因此它被标记为包括在过滤后的数据沿袭中。注意，第四数据变换节点116包括在嵌套于第一容器节点140中的第二容器节点142中。第一容器节点140和第二容器节点142均不与任何沿袭标签标识符相关联，因此它们俩也不包括在黑名单中。由于第二沿袭路径遍历了第一容器节点140和第二容器节点142的边界，并且两个容器节点都不在黑名单中，所以第一容器节点140和第二容器节点142都被标记为包括在过滤后的数据沿袭中。注意，在一些示例中，如果容器节点标记有沿袭标签标识符和值，则其所有后代节点(即，子节点及它们的孩子)都标记有相同的沿袭标签标识符和值。因此，黑名单可以基于最低级别的节点来构建。

[0088] 接下来，数据沿袭模块25遍历第四数据变换节点116和第一数据变换节点110之间的有向链路108。当沿袭模块25到达第一数据变换节点110时，搜索黑名单以确定黑名单是否包括第一数据变换节点110。由于第一数据变换节点110不在黑名单中，因此它被标记为包括在过滤后的数据沿袭中。由于第一数据变换节点110处于端到端沿袭图100的边缘，所以第二沿袭路径的遍历完成。

[0089] 参考图7，数据沿袭模块25遍历在第一数据节点136和第一数据变换节点110之间延伸的第三沿袭路径。同样，由于先前已经遍历了第一数据节点136和第十数据变换节点128之间的有向链路108，所以数据沿袭模块25从遍历第十数据变换节点128和第六数据变换节点120之间的有向链路108开始其遍历。

[0090] 当沿袭模块25到达第六数据变换节点120时，搜索黑名单以确定黑名单是否包括第六数据变换节点120。由于第六数据变换节点120不在黑名单中，因此它被标记为包括在过滤后的数据沿袭中。注意，第六数据变换节点120包括在嵌套于第一容器节点140中的第三容器节点144中。第一容器节点140和第三容器节点144都不与任何沿袭标签标识符相关联，因此它们均不包括在黑名单中。由于第二沿袭路径遍历了第一容器节点140和第三容器节点144的边界，并且两个容器节点均不在黑名单中，所以第一容器节点140和第三容器节点144均标记为包括在过滤后的数据沿袭中。

[0091] 接下来，数据沿袭模块25遍历第六数据变换节点120和第一数据变换节点110之间的有向链路108。当沿袭模块25到达第一数据变换节点110时，搜索黑名单以确定黑名单是否包括第一数据变换节点110。由于第一数据变换节点110不在黑名单中，因此它被标记为

包括在过滤后的数据沿袭中。由于第一数据变换节点110位于沿袭图100的边缘，所以第三沿袭路径的遍历完成。

[0092] 参考图8，数据沿袭模块25遍历在第一数据节点136和第二数据变换节点112之间延伸的第四沿袭路径。由于先前已经遍历了第一数据节点136和第十数据变换节点128之间的有向链路108，所以数据沿袭模块25从遍历第十数据变换节点128和第五数据变换节点118之间的有向链路108来开始其遍历。

[0093] 当沿袭模块25到达第五数据变换节点118时，搜索黑名单以确定黑名单是否包括第五数据变换节点118。由于第五数据变换节点118存在于黑名单中，所以从过滤后的数据沿袭中将它排除。由于第五数据变换节点118被确定为在黑名单中，所以终止第四沿袭路径的遍历。

[0094] 参考图9，数据沿袭模块25遍历在第一数据节点136和第三数据变换节点114之间延伸的第五沿袭路径。由于先前已经遍历了第一数据节点136和第十数据变换节点128之间的有向链路108，所以数据沿袭模块25从遍历第十数据变换节点128和第八数据变换节点124之间的有向链路108开始其遍历。

[0095] 当沿袭模块25到达第八数据变换节点124时，搜索黑名单以确定黑名单是否包括第八数据变换节点124。由于第八数据变换节点124不在黑名单中，所以它被标记为包括在过滤后的数据沿袭中。注意，第八数据变换节点124包括在嵌套于第一容器节点140中的第四容器节点146中。第一容器节点140和第四容器节点146均不与任何沿袭标签标识符相关联，因此，它们均不包括在黑名单中。由于第五沿袭路径遍历了第一容器节点140和第四容器节点146的边界，并且两个容器节点都不在黑名单中，所以第一容器节点140和第四容器节点146均标记为包括在过滤后的数据沿袭中。

[0096] 接下来，数据沿袭模块25遍历第八数据变换节点124和第三数据变换节点114之间的有向链路。当沿袭模块25到达第三数据变换节点114时，搜索黑名单以确定黑名单是否包括第三数据变换节点114。由于第三数据变换节点114不在黑名单中，因此它被标记为包括在过滤后的数据沿袭中。由于第三数据变换节点114位于沿袭图100的边缘，所以第五沿袭路径的遍历完成。

[0097] 参考图10，数据沿袭模块25遍历在第一数据节点136和第二数据变换节点112之间延伸的第六沿袭路径。由于先前已经遍历了第一数据节点136和第十数据变换节点128之间的有向链路108，所以数据沿袭模块25从遍历第十数据变换节点128和第十一数据变换节点130之间的有向链路108开始其遍历。

[0098] 当沿袭模块25到达第十一数据变换节点130时，搜索黑名单以确定黑名单是否包括第十一数据变换节点130。由于第十一数据变换节点130不在黑名单中，所以它被标记为包括在过滤后的数据沿袭中。

[0099] 接下来，数据沿袭模块25遍历第十一数据变换节点130和第七数据变换节点122之间的有向链路。当沿袭模块25到达第七数据变换节点122时，搜索黑名单以确定黑名单是否包括第七数据变换节点122。由于第七数据变换节点122存在于黑名单中，所以从过滤后的数据沿袭中将它排除。由于第七数据变换节点122被确定为存在于黑名单中，所以终止第六沿袭路径的遍历。

[0100] 参考图11，由于第十二数据变换组件132、第十三数据变换组件134和第二数据节

点138不在从第一数据节点136开始的沿袭路径上，所以默认将它们从过滤后的数据沿袭中排除。

[0101] 注意，在任意有向链路的遍历过程中，从来没有到达过第二数据变换节点112，因此从过滤后的数据沿袭中将它排除。

[0102] 参考图12，一旦遍历完从第一数据节点136开始的所有沿袭路径，则从端到端数据沿袭图100中移除未被标记为包括在过滤后的数据沿袭中的任何节点，从而得到过滤后的数据沿袭图200。注意，过滤后的数据沿袭图200不包括与“区域”沿袭标签标识符相关联且具有仅包括“CA”的沿袭标签值的节点。

[0103] 2标准化/去标准化

[0104] 参考图13，计算机程序27的一个简单示例性的端到端数据沿袭图1300包括输入表单1360、标准化数据变换节点1362、数据节点1364、去标准化数据变换节点1366和报告表单1368。输入表单1360和报告表单1368是多个数据节点，这些数据节点中的每一个数据节点均包括这些数据节点中的与嵌套在较高级数据元素内的不同数据元素对应的多个数据节点。在此示例中，嵌套的数据元素表示每个表单记录中的不同字段。

[0105] 标准化数据变换节点1362执行标准化操作，该标准化操作对存储在由数据节点1354表示的数据存储器中的数据元素进行标准化，去标准化数据变换节点1366执行去标准化操作，该去标准化操作对从由数据节点1354表示的数据存储器中取回的数据元素进行去标准化。虽然标准化操作的“标准化”与去标准化操作的“去标准化”的具体实现方式可能不同，但通常它们具有以下特征。对数据元素进行标准化包括将对应于多个数据元素的记录转换成具有针对至少一个字段的通用格式的多个记录。对数据元素进行去标准化包括将具有针对至少一个字段的通用格式的多个记录转换成对应于多个数据元素的单个记录。以下是在数据沿袭图1300的上下文中的这种操作的示例。

[0106] 输入表单1360包括包含移动电话号码的第一字段1370、包含工作电话号码的第二字段1372和包含家庭电话号码的第三字段1374。在操作中，标准化数据变换节点1362接收输入表单1360，并将输入表单1360中所包括的电话号码转换成适于存储在数据节点1364中的标准化形式。在一些示例中，标准化数据变换节点1362将输入表单1360中所包括的电话号码转换成[电话号码类型，电话号码]形式，其中电话号码是具有三个值(移动、工作和家庭)的枚举数据类型。在本文所述方案的上下文中，将具有数据元素类型和数据元素值的数据元素转换成包括数据元素类型字段和数据元素值字段的数据结构的上述过程被称为对数据元素进行“标准化”。标准化后的电话号码存储在数据节点1364中以供后续使用。

[0107] 去标准化数据变换节点1366从数据节点1364中读取标准化后的电话号码，并对标准化后的电话号码进行去标准化以使其还原成原始形式。在本文所述方案的上下文中，将标准化数据结构转换回具有数据元素类型和数据元素值的数据元素的上述过程被称为对数据元素进行“去标准化”。去标准化数据变换节点1366接下来将移动电话号码写入用于存储移动电话号码的报告表单1368的第四字段1376中，用于存储工作电话号码的报告表单1368的第五字段1378中和用于存储家庭电话号码的报告表单1368的第六字段1380中。

[0108] 因为从数据项成为标准化枚举数据类型的转换以及从标准化枚举数据类型成为数据项的转换，所以对于常规数据沿袭计算系统来说，上述标准化/去标准化处理可能是困难的。例如，如果用户请求报告表单1368第四字段1376中的移动电话号码的数据沿袭图，则

依传统生成的数据沿袭图将包括报告表单1368第四字段1376中的移动电话号码、去标准化数据变换节点1366、数据节点1364、标准化数据变换节点1362和输入表单1368的第一字段1370。

[0109] 依传统生成的数据沿袭图还包括输入表单1360第二字段1372中的工作电话号码以及输入表单1360第三字段1374中的家庭电话号码。这样的传统数据沿袭图是不正确的，因为第二字段1372中的工作电话号码和第三字段1374中的家庭电话号码实际上不是报告表单1368第四字段1376中的移动电话号码的数据沿袭的一部分。

[0110] 参考图14，为了防止生成不正确的数据沿袭图，端对端数据沿袭图1300中的各个字段(其本质上是数据节点)标注沿袭标签，该沿袭标签具有沿袭标签标识符“电话号码”和可能的沿袭标签值：“移动”、“工作”和“家庭”。具体地，第一字段1370标注有沿袭标签，该沿袭标签具有“电话号码”沿袭标签标识符和仅包括“移动”的可能沿袭标签值的子集。第二字段1372标注有沿袭标签，该沿袭标签具有“电话号码”沿袭标签标识符和仅包括“工作”的可能沿袭标签值的子集。第三字段1374标注有沿袭标签，该沿袭标签具有“电话号码”沿袭标签标识符和仅包括“家庭”的可能沿袭标签值的子集。

[0111] 第四字段1376标注有沿袭标签，该沿袭标签具有“电话号码”沿袭标签标识符和仅包括“移动”的可能沿袭标签值的子集。第五字段1378标注有沿袭标签，该沿袭标签具有“电话号码”沿袭标签标识符和仅包括“工作”的可能沿袭标签值的子集。第六个字段1380签注有沿袭标签，该沿袭标签具有“电话号码”沿袭标签标识符和仅包括“家庭”的可能的沿袭标签值的子集。

[0112] 参考图15，当用户选择包括移动电话号码的第四字段1376以生成过滤后的数据沿袭图时，数据沿袭模块25首先确定第四字段1376与一个或多个沿袭标签标识符之间的关联性。在这种情况下，第四字段1376与“电话号码”沿袭标签标识符相关联。接下来，数据沿袭模块25确定“电话号码”标签标识符的可能标签值中的哪个标签值与第四字段1376相关联。在这种情况下，“移动”标签值是与第四字段1376相关联的唯一标签值。

[0113] 接下来，数据沿袭模块25处理端到端数据沿袭图1300以生成黑名单。如上所述，检查除所选第四字段1376之外的端到端数据沿袭图1300中的所有节点和字段来生成黑名单，以便识别与“电话号码”沿袭标签标识符相关联且具有仅包括“工作”和/或“家庭”的标签值子集的节点和字段。将图3的过滤算法300应用到端到端数据沿袭图1300导致第二字段1372、第三字段1374、第五字段1378和第六字段1380被添加到黑名单。在图15中，用对角线图案填充被添加到黑名单的数据沿袭字段。

[0114] 参考图16，接下来数据沿袭模块25遍历从第四字段1376开始的所有沿袭路径，以确定哪些节点和字段被包括在过滤后的数据沿袭中以及从过滤后的数据沿袭中排除哪些节点和字段。从图中可以看出，数据沿袭模块25将仅用“工作”和“家庭”沿袭标签值标注的字段标记为从过滤后的数据沿袭中排除，而将仅用“移动”沿袭标签值标注的字段标记为包括在过滤后的数据沿袭中。

[0115] 参考图17，一旦遍历了从第四字段1376开始的所有沿袭路径，则从端到端数据沿袭图1300中移除未被标记为包括在过滤后的数据沿袭中的任何节点和字段，这就形成了过滤后的数据沿袭图1700。注意，从过滤后的数据沿袭图1700中排除了通过常规的沿袭生成技术而可能错误地包括在过滤后的数据沿袭图中的第二字段1327和第三字段1374。

[0116] 3替代方案

[0117] 应当注意，虽然以特定顺序遍历了上述基于沿袭标签的过滤示例中的数据沿袭图的节点，但是可以使用许多不同的遍历顺序。例如，可以使用广度优先遍历顺序(breadth first order)或深度优先遍历顺序(depth first order)。

[0118] 在上述基于沿袭标签的过滤示例中，选择初始节点，并且使用与初始节点相关联的沿袭标签和沿袭标签值来过滤沿袭图。然而，在一些示例中，用户选择初始节点，并且数据沿袭模块遍历数据沿袭图以确定初始节点的完整数据沿袭，并且发现可用于过滤该完整数据沿袭的任何沿袭标签和沿袭标签值。然后将初始节点的完整数据沿袭连同所发现的沿袭标签和沿袭标签值一起呈现给用户。接下来，用户可以通过选择一个或多个所发现的沿袭标签和沿袭标签值作为过滤参数来过滤该完整数据沿袭。在其他示例中，在数据沿袭模块遍历数据沿袭图时，维护所遇到的沿袭标签和沿袭标签值的列表。然后，可以在遍历图的第二上游部分时使用在遍历图的第一部分时被添加至列表中的一些或所有沿袭标签和沿袭标签值，以过滤图的第二部分。

[0119] 虽然上述方法通常与数据沿袭图的生成相关，但它们也可用于生成影响分析图。

[0120] 在一些示例中，与处于沿袭层级中的不同层(例如，容器节点和容器节点内的数据变换节点)的两个或更多个节点相关联的沿袭标签可能会发生冲突。在一些示例中，选择处于沿袭层级中最底层的节点作为沿袭标签。

[0121] 在上述示例中，数据变换节点被示出为通过以箭头表示的流而彼此直接连接。然而，应当注意，在一些示例中，将第一组件的输出连接到第二组件的输入的流可以隐含地包括数据存储元件，该数据存储元素被从第一组件出来的数据写入，并被第二个组件的输入读取。

[0122] 4实施方式

[0123] 上述数据沿袭生成方法可以例如使用执行合适的软件的计算系统来实现。例如，软件可以包括在一个或多个编程或可编程计算系统(其可以是诸如分布式、客户端/服务器或网格的各种架构)上执行的一个或多个计算机程序中的过程，每个计算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/或非易失性存储器和/或存储元件)、至少一个用户界面(用于使用至少一个输入设备或端口接收输入，并且用于使用至少一个输出设备或端口提供输出)。软件可以包括例如提供与数据流图的设计、配置和执行相关的服务的较大程序的一个或多个模块。程序的模块(例如，数据流图的元素)可以被实现为符合存储在数据仓库中的数据模型的数据结构或其它组织形式的数据。

[0124] 软件可以提供在有形、非暂时性介质上，例如CD-ROM或其他计算机可读介质(例如，可由通用或专用计算系统或设备读取)，或者通过网络的通信介质被递送(例如，被编码成传播信号)到其被执行的计算系统的有形、非暂时性介质。可以在专用计算机上或使用诸如协处理器或现场可编程门阵列(FPGA)或特定的专用集成电路(ASIC)之类的专用硬件来执行一些或全部处理。处理可以以分布式方式实现，其中由软件指定的不同计算部分由不同的计算元件执行。每个这样的计算机程序优选地存储在或下载到可由通用或专用可编程计算机访问的存储设备的计算机可读存储介质(例如，固态存储器或媒介，或磁媒介或光媒介)上，用于当计算机读取存储设备介质以执行本文所述的处理时，配置和操作计算机。本发明的系统还可以被认为可实现为配置有计算机程序的有形的、非暂时性介质，其中如此

配置的介质使得计算机以特定和预定义的方式操作以执行本文描述的一个或多个处理步骤。

[0125] 已经描述了本发明的多个实施例。然而，应当理解，前述描述旨在说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的范围限定。因此，其他实施例也在所附权利要求的范围内。例如，在不脱离本发明的范围的情况下可以进行各种修改。另外，上述的一些步骤可以是与顺序无关的，并且因此可以以与所描述的顺序不同的顺序来执行。

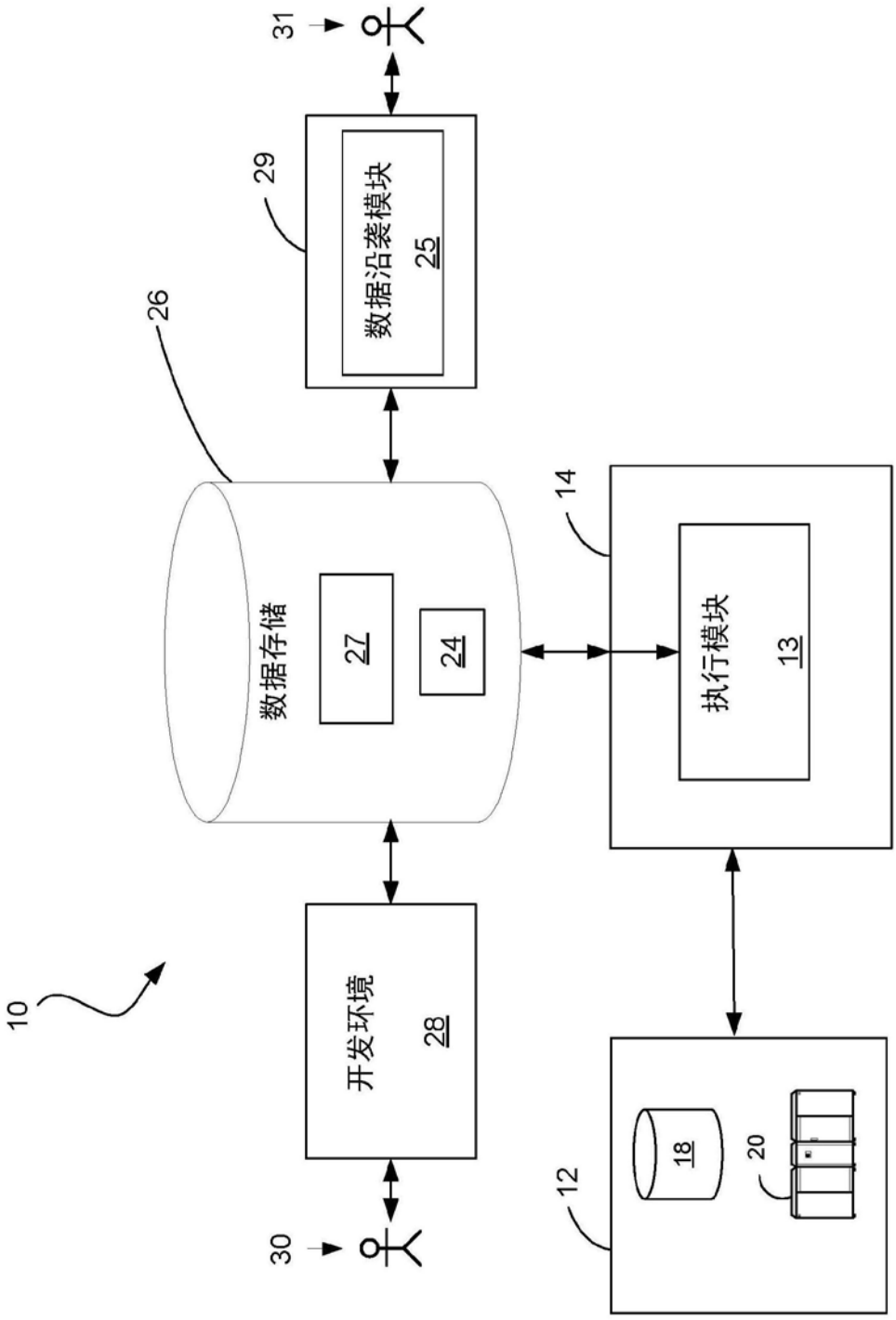


图1

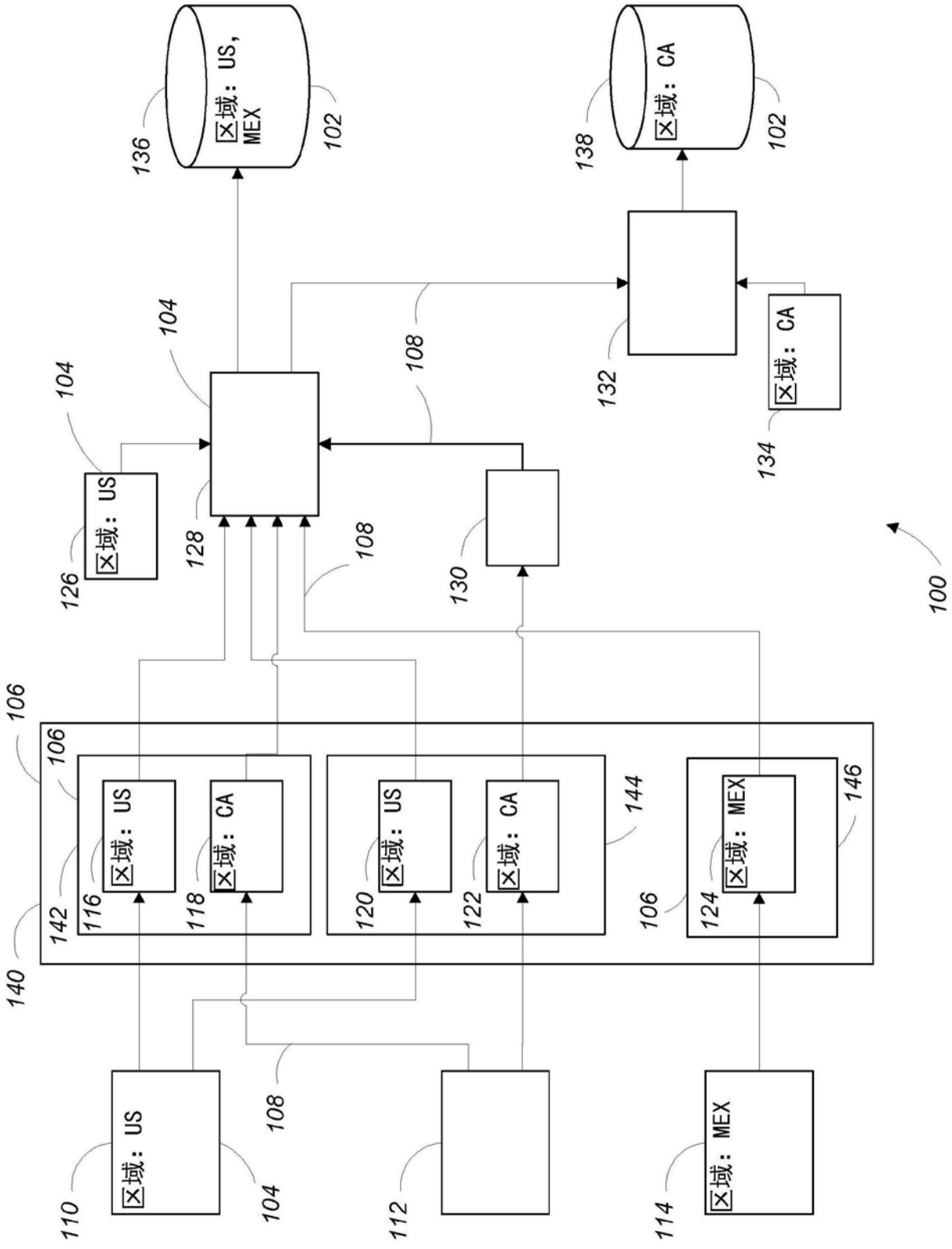

区域：US，MEX
区域：CA
区域：US
区域：CA
区域：US
区域：CA
区域：MEX
区域：US
区域：MEX
100
102
104
106
108
110
112
114
116
118
120
122
124
126
128
130
132
134
136
138
140
142
144
146

图2a

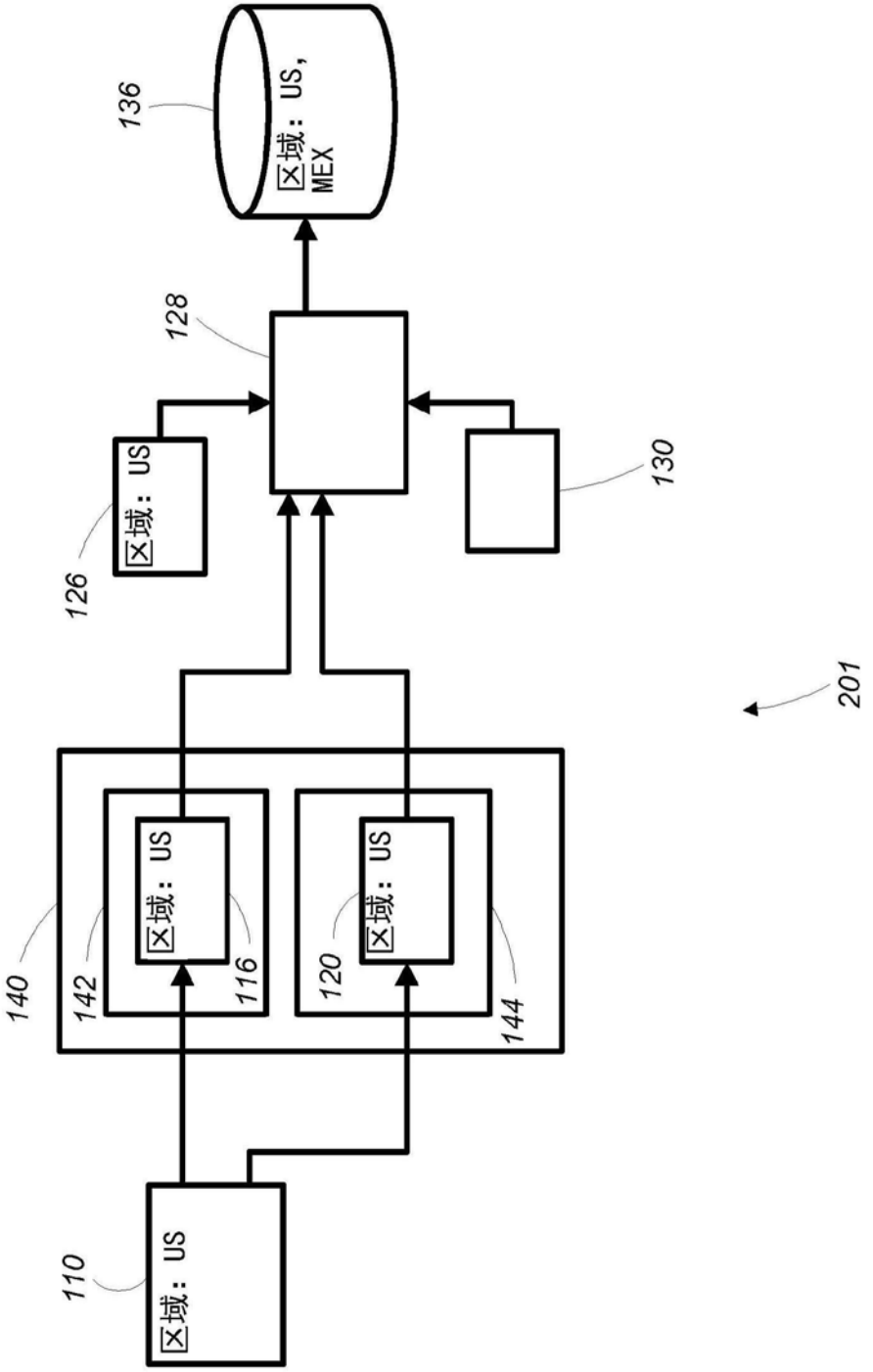

136
区域：US，MEX
128
126
区域：US
130
201
140
142
区域：US
116
120
区域：US
144
110
区域：US

图2b

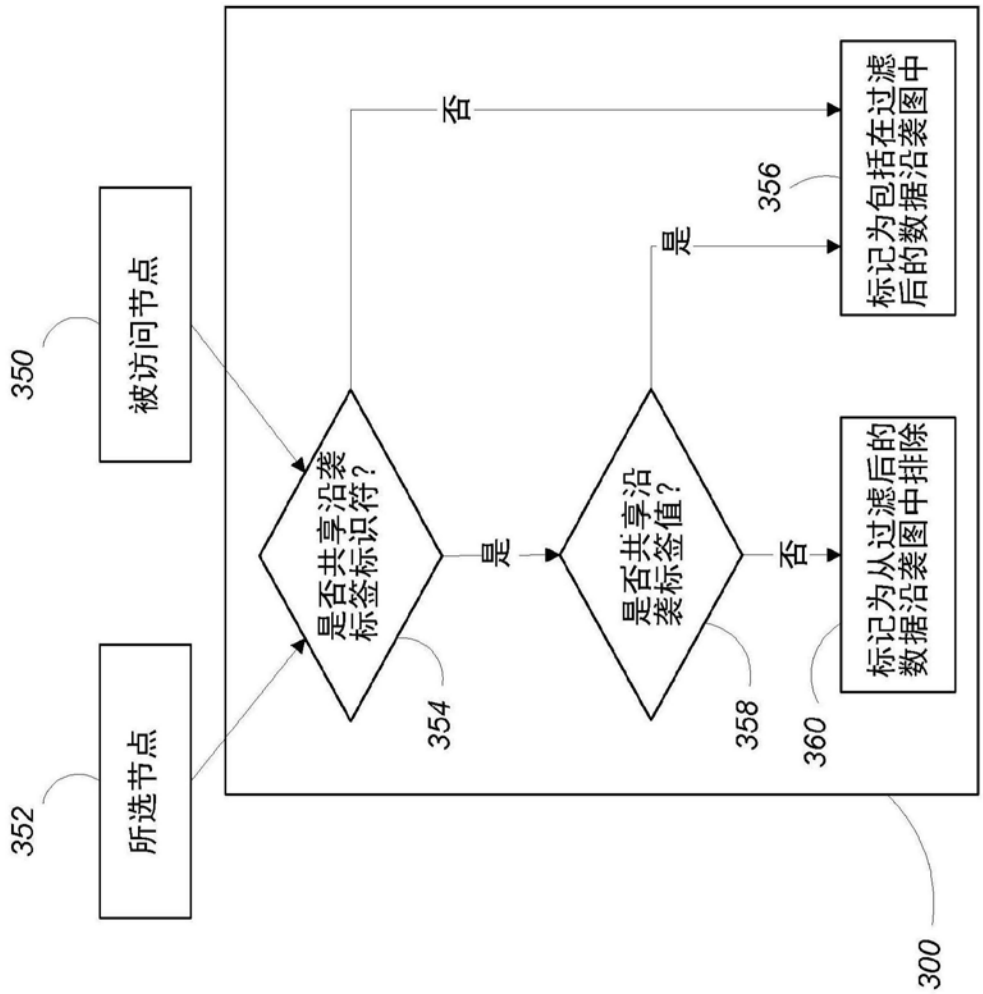

被访问节点
350
所选节点
352
是否共享沿袭标识符？
354
是
是否共享沿袭标签值？
358
否
标记为从过滤后的数据沿袭图中排除
360
否
是
标记为包括在过滤后的数据沿袭图中
356
300

图3

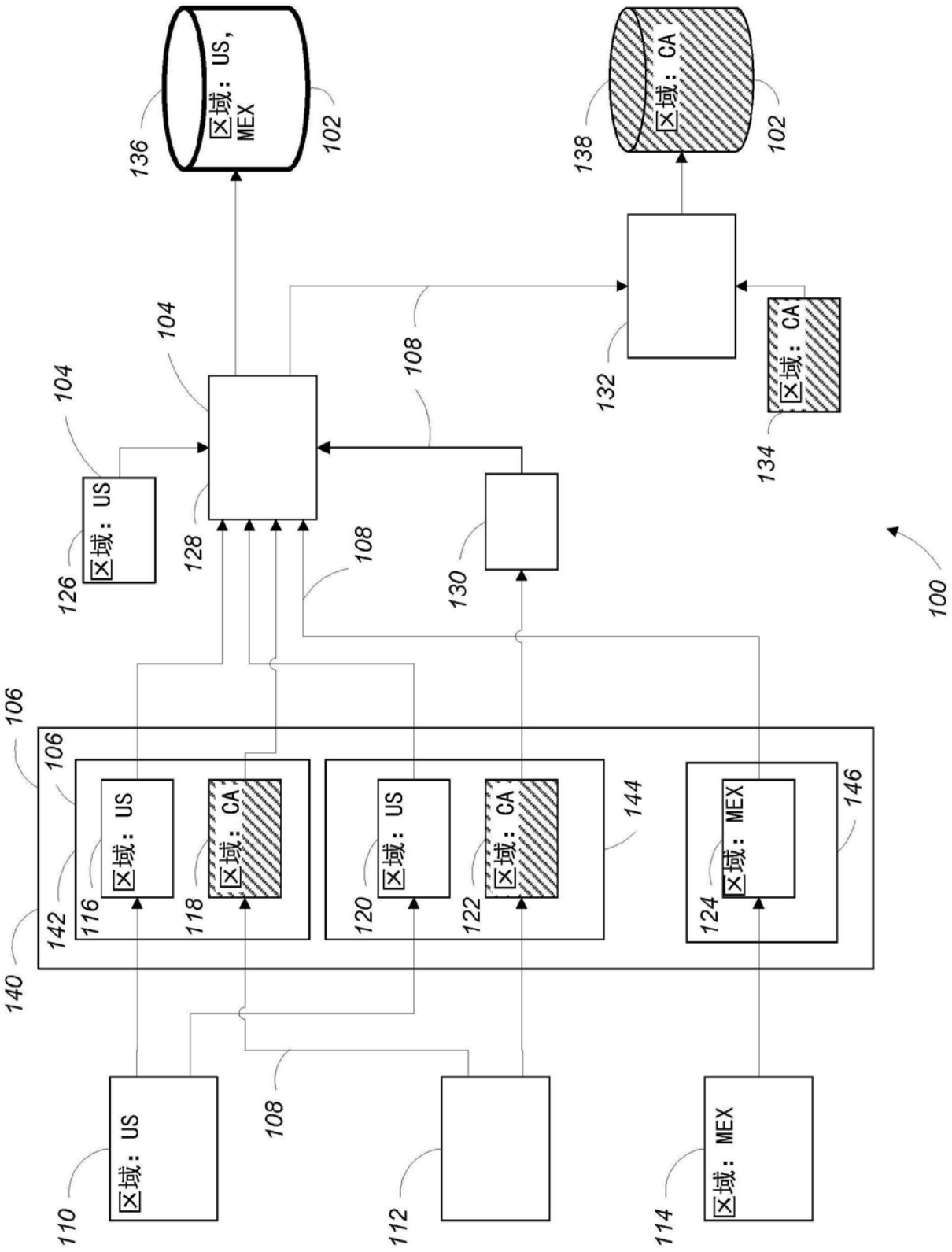

区域：US，MEX
区域：CA
区域：US
区域：CA
区域：US
区域：CA
区域：MEX
区域：US
区域：MEX
100
102
102
104
104
106
106
108
108
108
110
112
114
116
118
120
122
124
126
128
130
132
134
136
138
140
142
144
146

图4

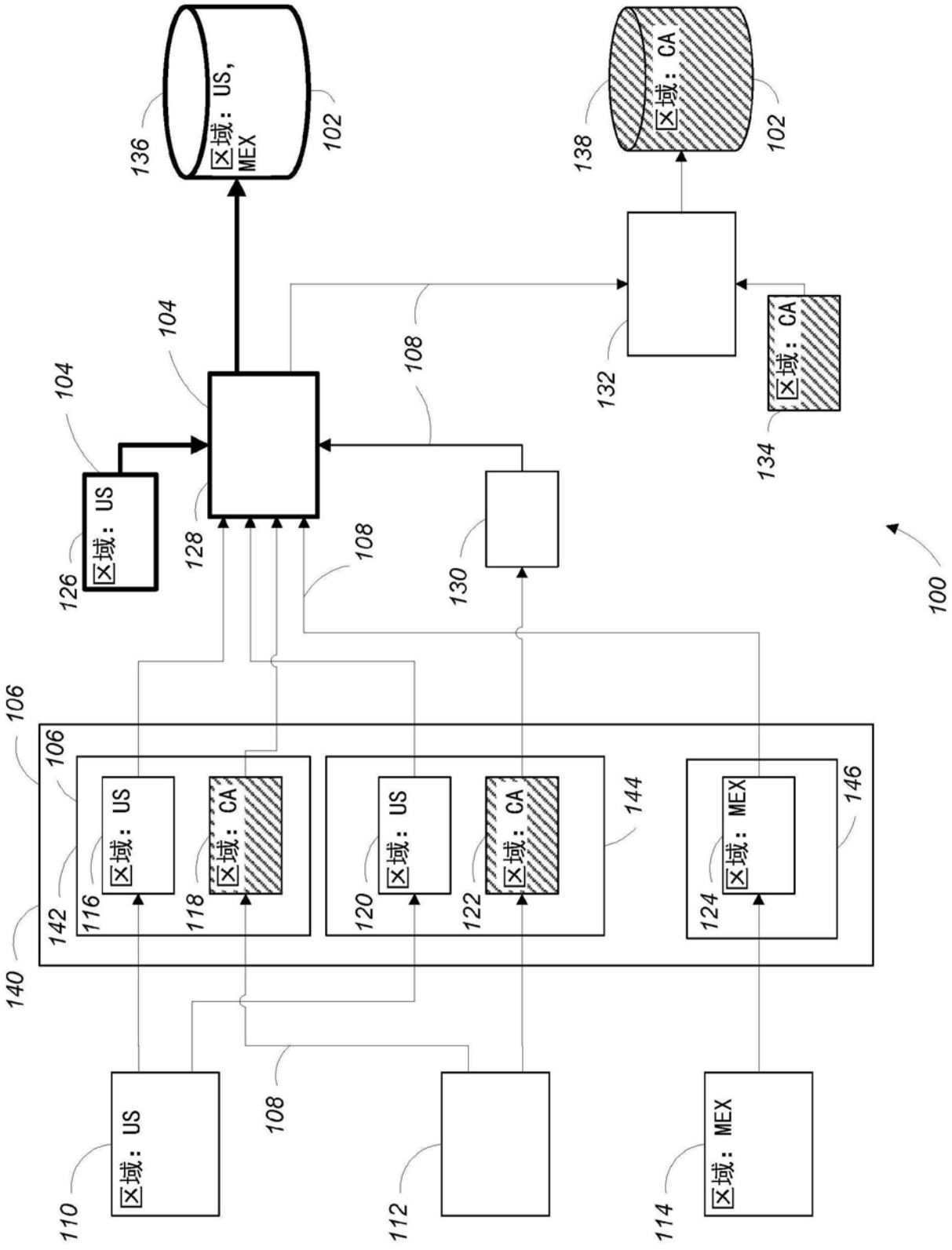

区域：US，MEX
区域：CA
区域：US
区域：CA
区域：US
区域：CA
区域：MEX
区域：US
区域：MEX
100
102
102
104
104
106
106
108
108
108
110
112
114
116
118
120
122
124
126
128
130
132
134
136
138
140
142
144
146

图5

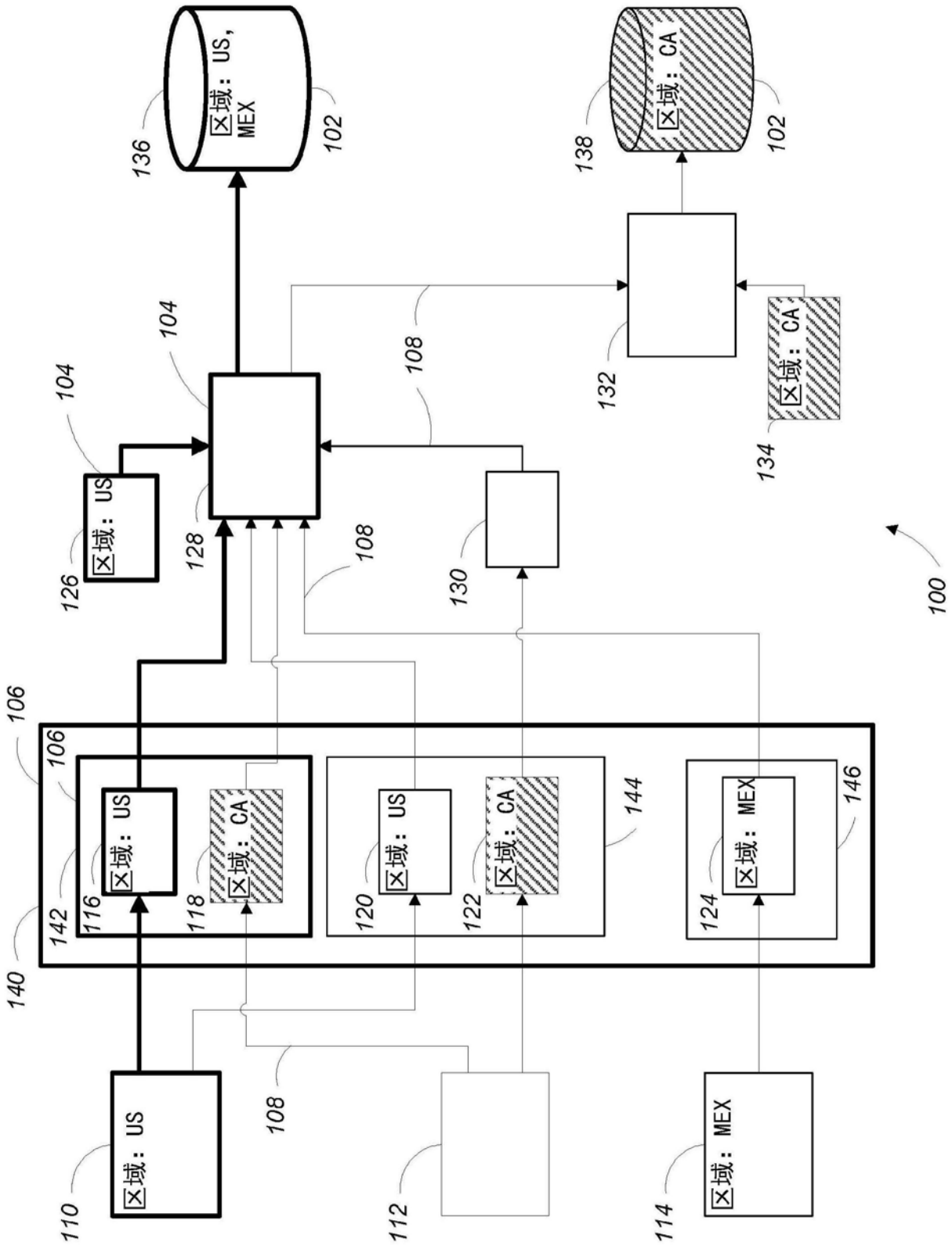

区域：US，MEX
区域：CA
区域：CA
区域：US
区域：US
区域：CA
区域：US
区域：CA
区域：MEX
区域：US
区域：MEX
100
102
102
104
104
106
106
108
108
108
110
112
114
116
118
120
122
124
126
128
130
132
134
136
138
140
142
144
146

图6

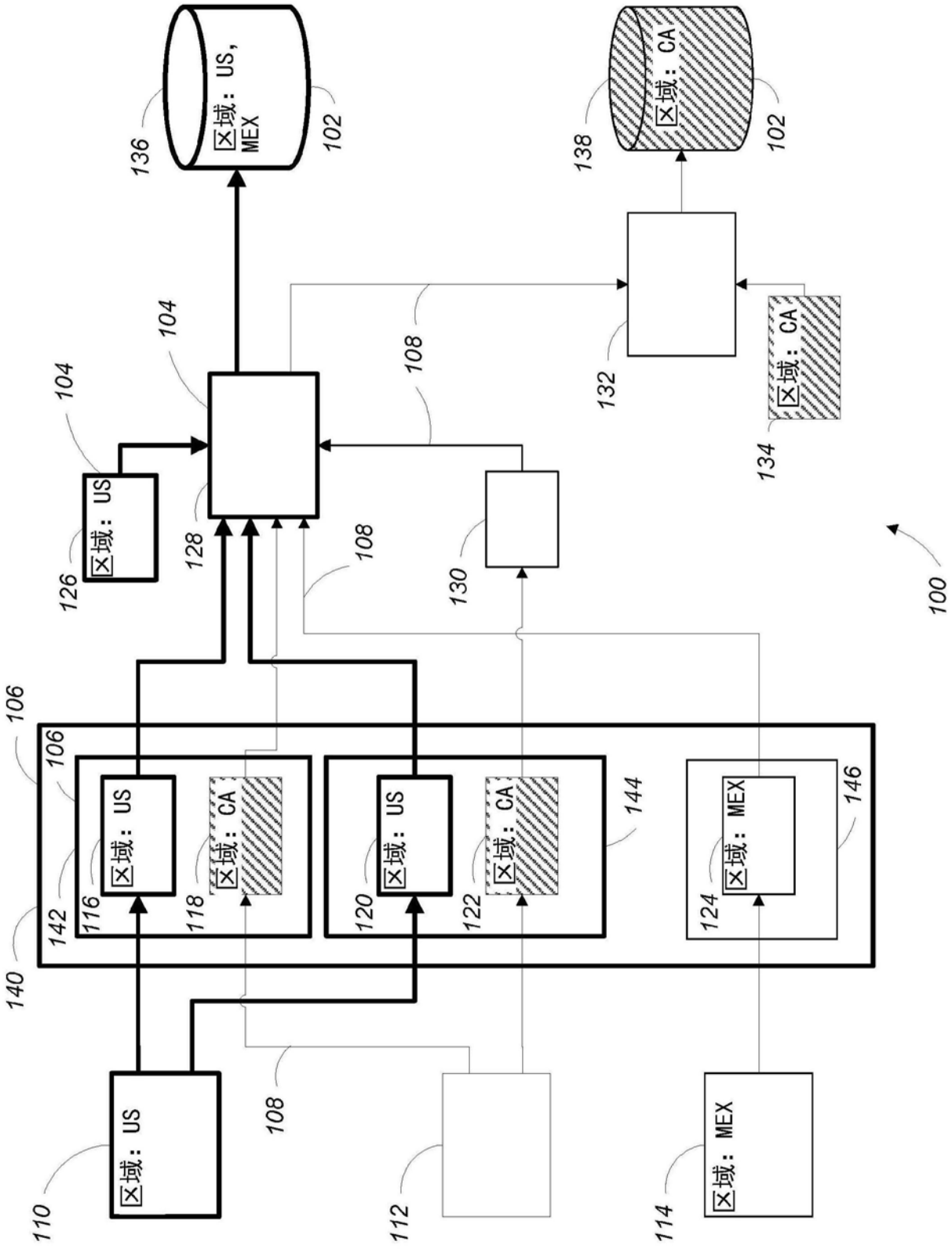

区域：US，MEX
区域：CA
区域：US
区域：CA
区域：US
区域：CA
区域：MEX
区域：US
区域：MEX
100
102
102
104
104
106
106
108
108
108
110
112
114
116
118
120
122
124
126
128
130
132
134
136
138
140
142
144
146

图7

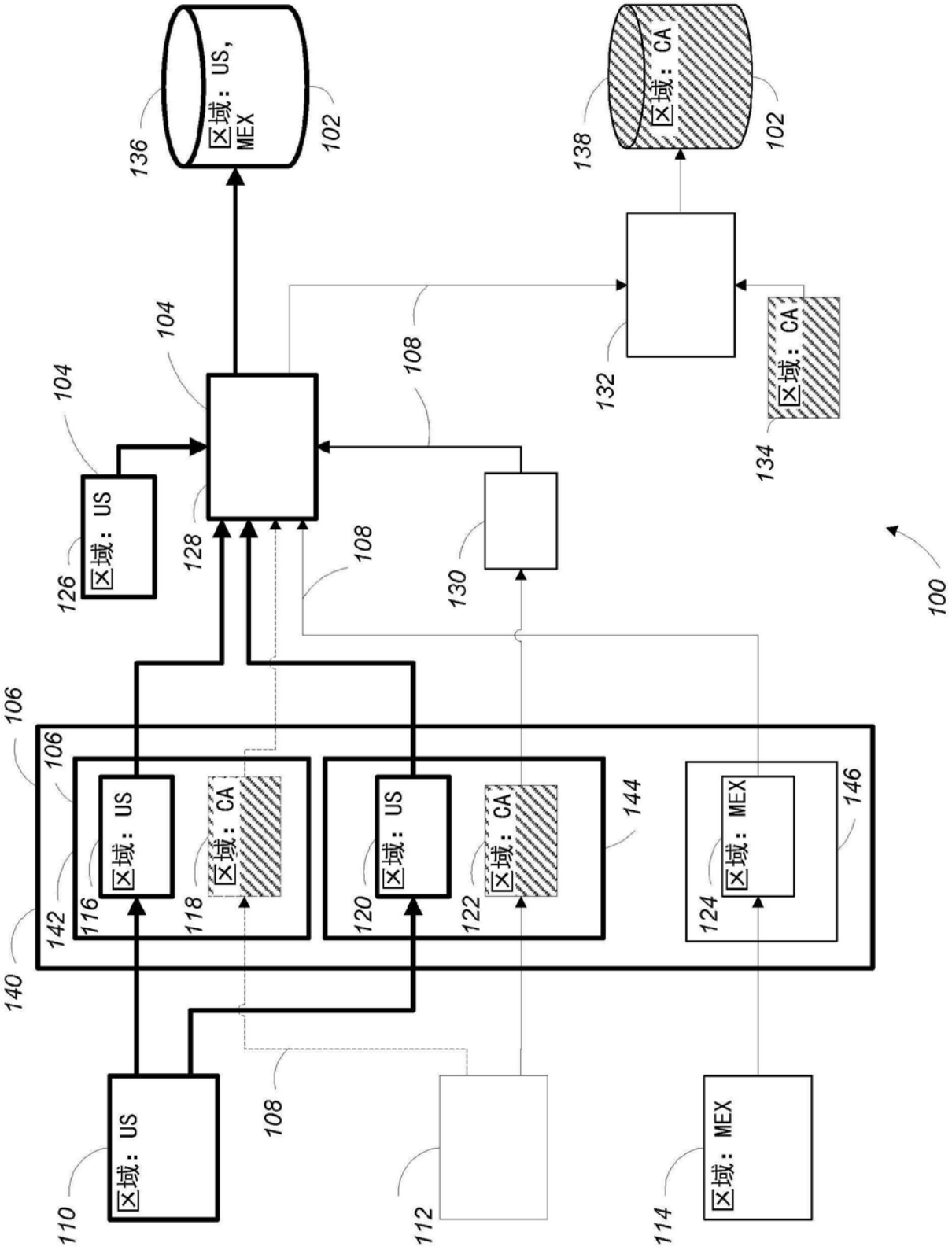

区域：US，MEX
区域：CA
区域：US
区域：CA
区域：US
区域：CA
区域：US
区域：MEX
区域：US
区域：MEX
100
102
102
104
104
106
106
108
108
108
110
112
114
116
118
120
122
124
126
128
130
132
134
136
138
140
142
144
146

图8

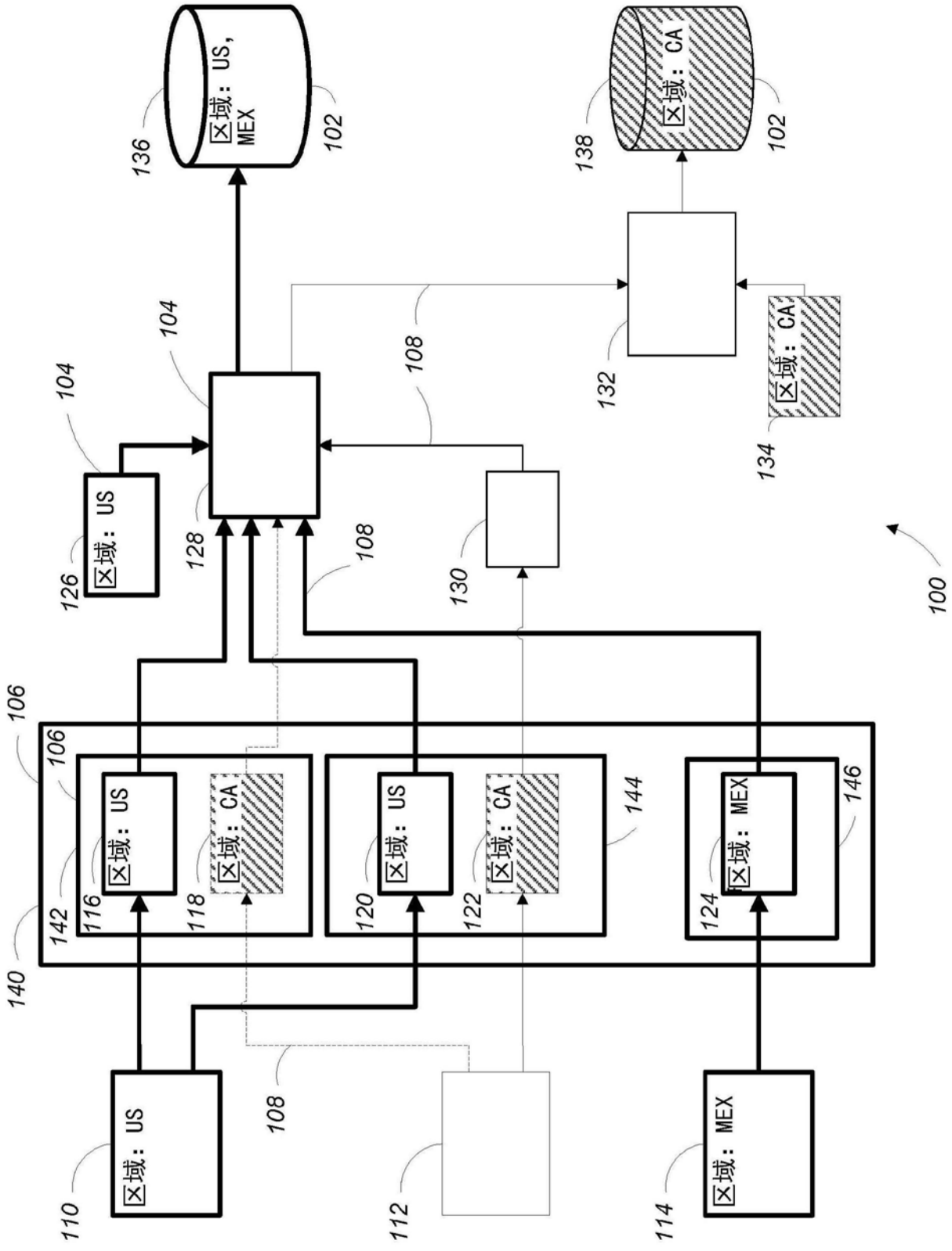

区域：US，MEX
区域：CA
区域：US
区域：CA
区域：US
区域：CA
区域：MEX
区域：US
区域：MEX
100
102
102
104
104
106
106
108
108
108
108
110
112
114
116
118
120
122
124
126
128
130
132
134
136
138
140
142
144
146

图9

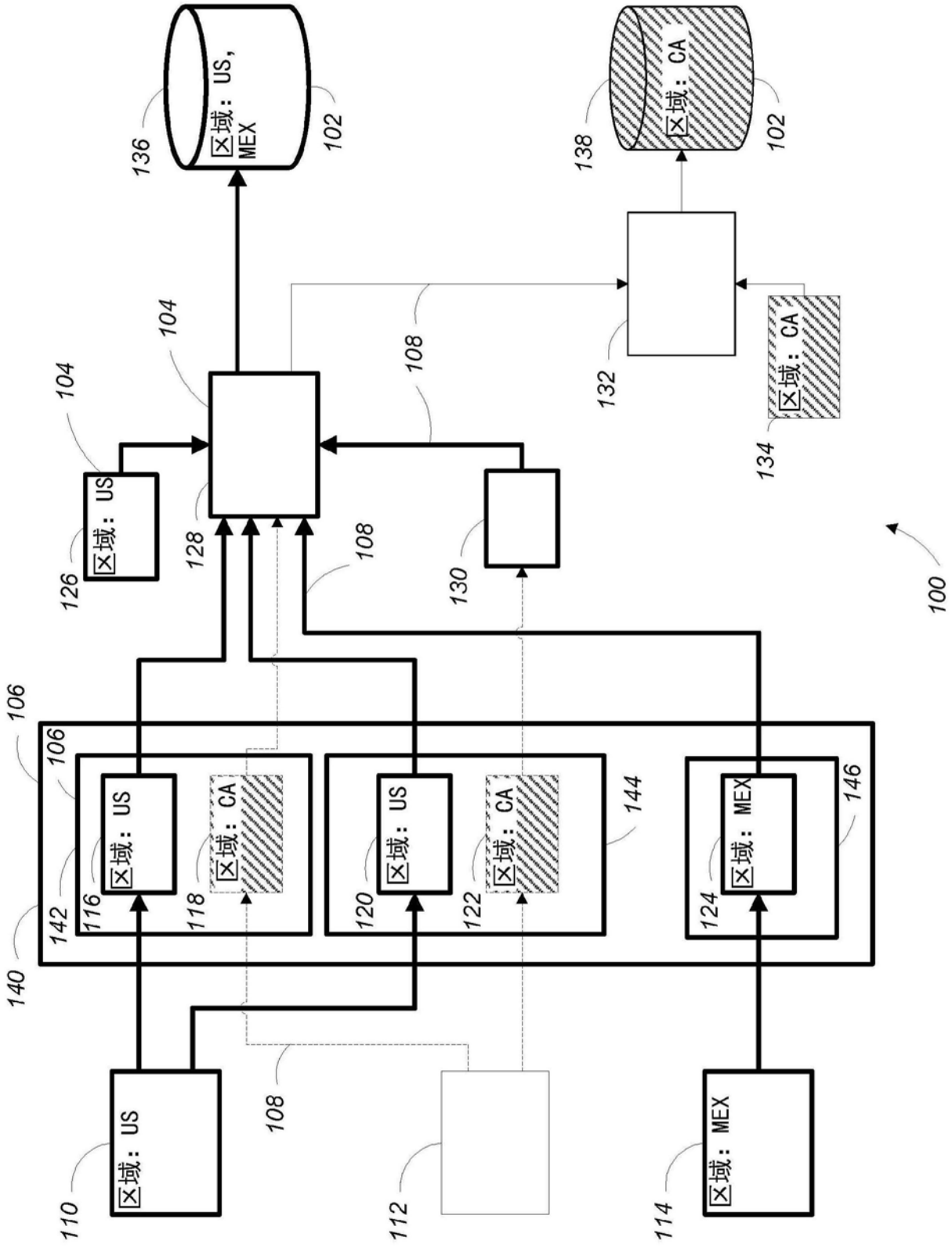

区域：US，MEX
区域：CA
区域：CA
区域：US
区域：US
区域：CA
区域：US
区域：CA
区域：MEX
区域：US
区域：MEX
100
102
102
104
104
106
106
108
108
108
110
112
114
116
118
120
122
124
126
128
130
132
134
136
138
140
142
144
146

图10

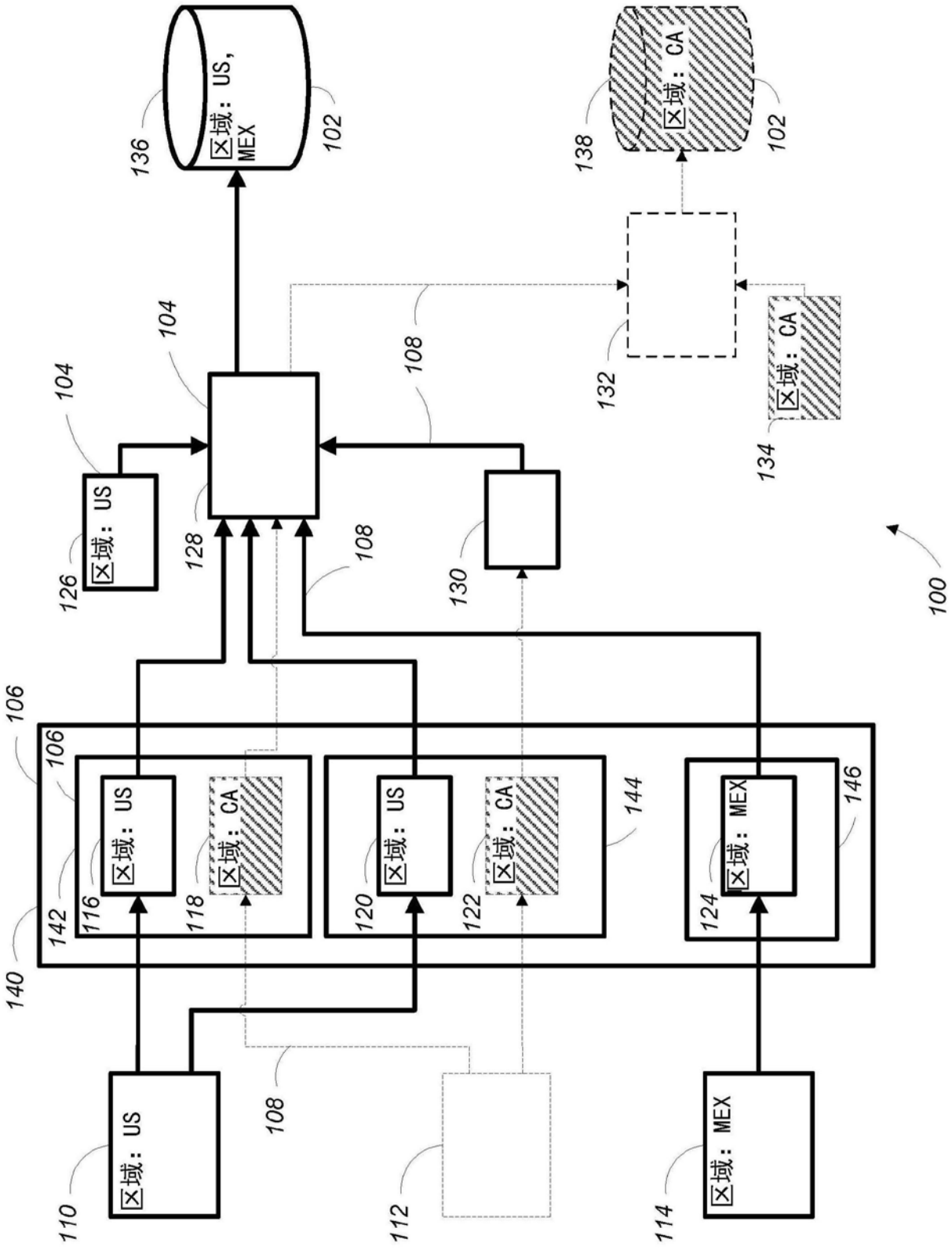

区域：US，MEX
区域：CA
区域：US
区域：CA
区域：US
区域：MEX
区域：US
区域：MEX
区域：US
区域：CA
区域：US
区域：CA
区域：US
区域：CA
区域：US
区域：MEX
100
102
102
104
104
106
106
108
108
108
108
110
112
114
116
118
120
122
124
126
128
130
132
134
136
138
140
142
144
146

图11

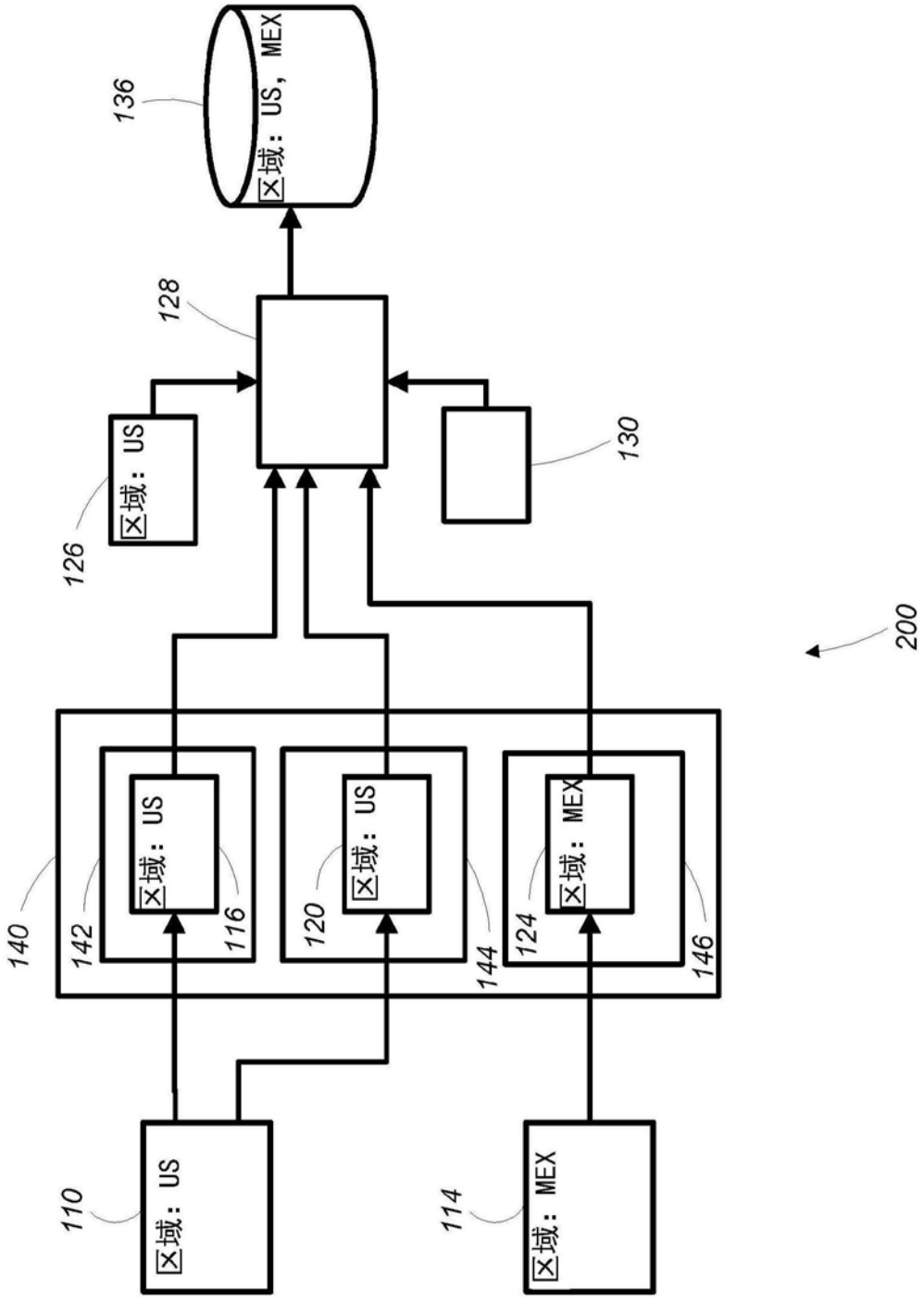

136
区域：US，MEX
128
126
区域：US
130
200
140
142
区域：US
116
120
区域：US
144
124
区域：MEX
146
110
区域：US
114
区域：MEX

图12

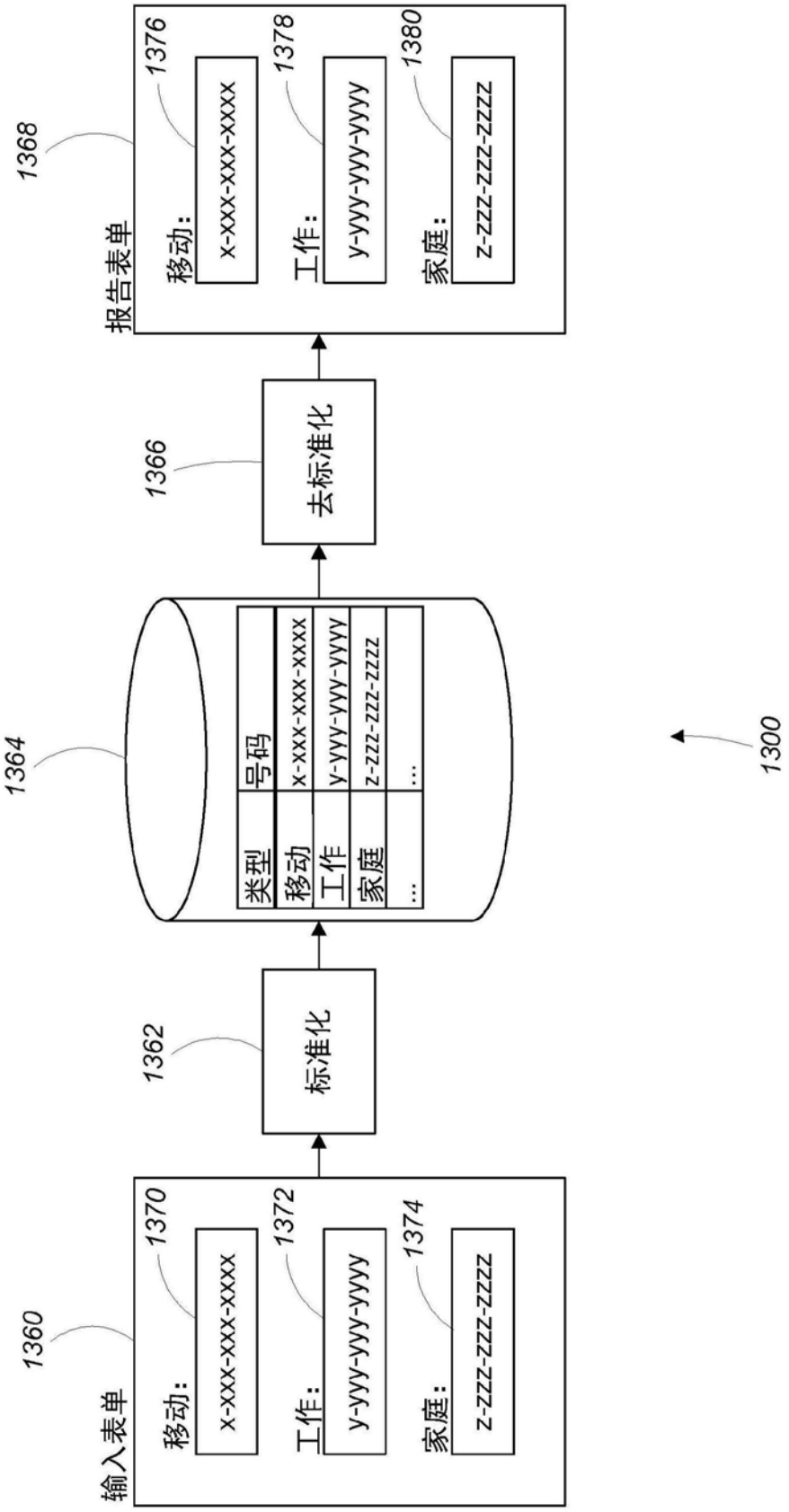

1360
输入表单
移动:
x-xxx-xxx-xxxx
工作:
y-yyy-yyy-yyyy
家庭:
z-zzz-zzz-zzzz
1370
1372
1374
1362
标准化
1364
类型
号码
移动
x-xxx-xxx-xxxx
工作
y-yyy-yyy-yyyy
家庭
z-zzz-zzz-zzzz
...
1366
去标准化
1368
报告表单
移动:
x-xxx-xxx-xxxx
工作:
y-yyy-yyy-yyyy
家庭:
z-zzz-zzz-zzzz
1376
1378
1380
1300

图13

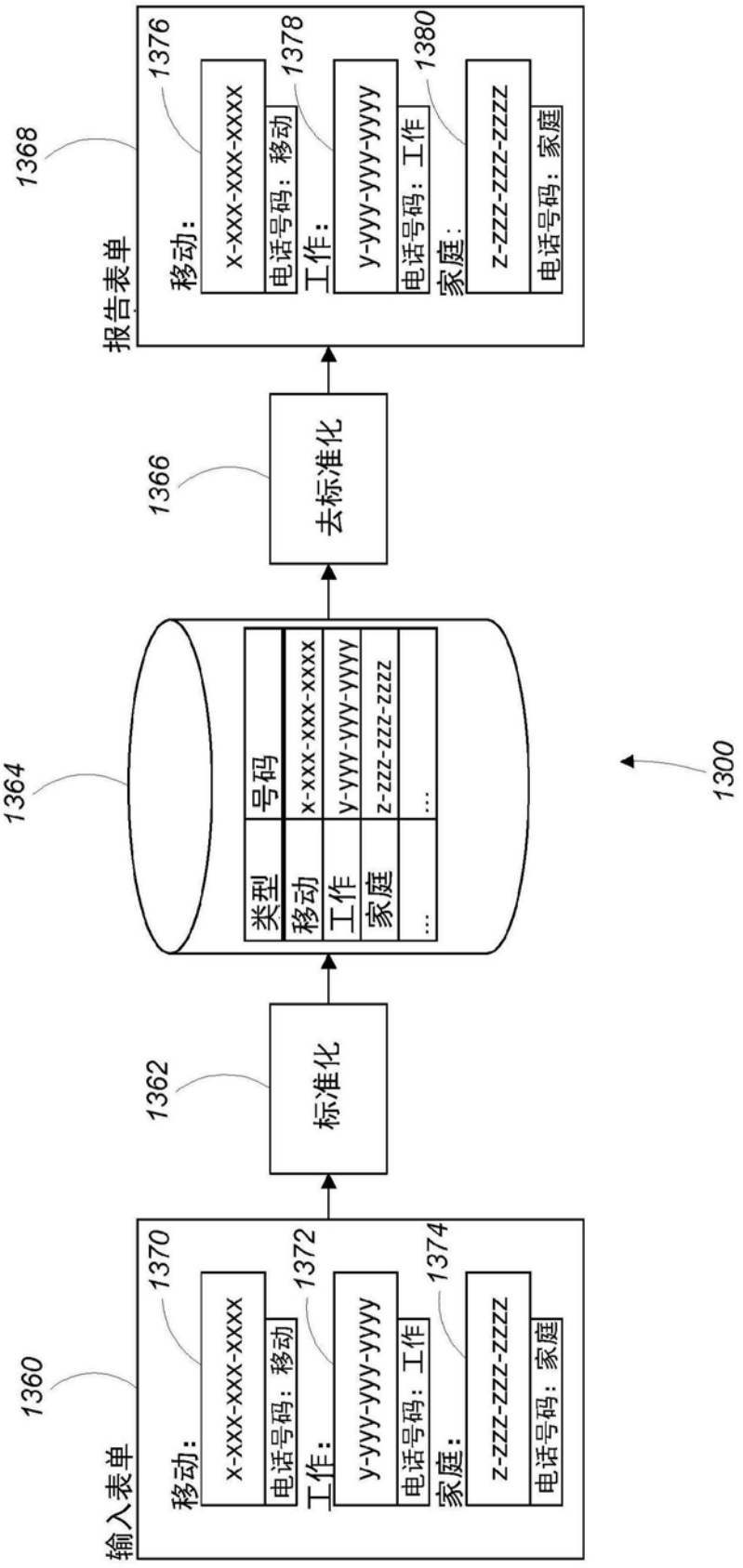

1360
输入表单
移动:
x-xxx-xxx-xxxx
电话号码：移动
工作:
y-yyy-yyy-yyyy
电话号码：工作
家庭:
z-zzz-zzz-zzzz
电话号码：家庭
1370
1372
1374
1362
标准化
1364
类型
号码
移动
x-xxx-xxx-xxxx
工作
y-yyy-yyy-yyyy
家庭
z-zzz-zzz-zzzz
...
...
1366
去标准化
1368
报告表单
移动:
x-xxx-xxx-xxxx
电话号码：移动
工作:
y-yyy-yyy-yyyy
电话号码：工作
家庭:
z-zzz-zzz-zzzz
电话号码：家庭
1376
1378
1380
1300

图14

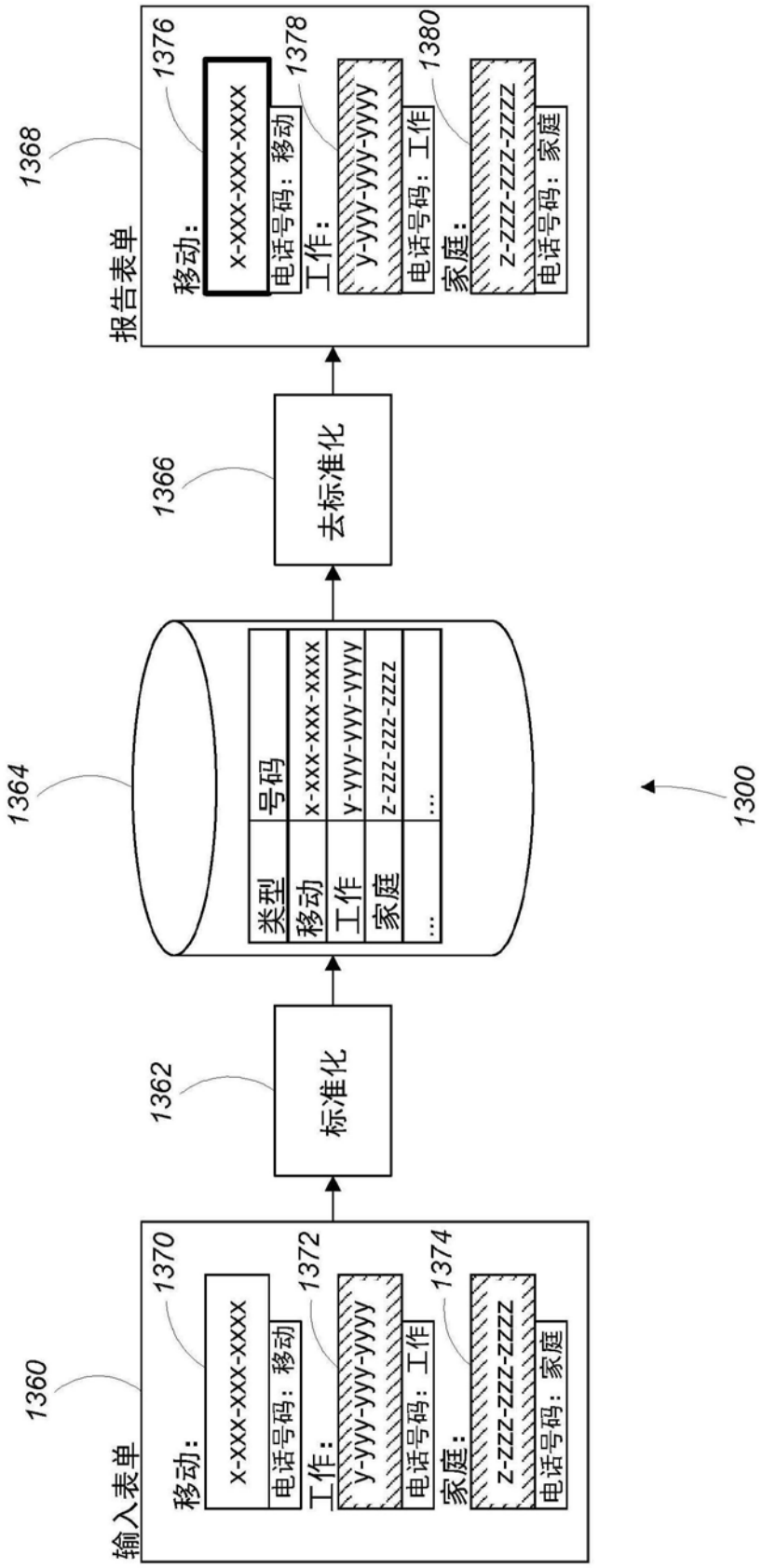

1360
输入表单
移动：
x-xxx-xxx-xxxx
电话号码：移动
工作：
y-yyy-yyy-yyyy
电话号码：工作
家庭：
z-zzz-zzz-zzzz
电话号码：家庭
1370
1372
1374
1362
标准化
1364
类型
号码
移动
x-xxx-xxx-xxxx
工作
y-yyy-yyy-yyyy
家庭
z-zzz-zzz-zzzz
...
1366
去标准化
1368
报告表单
移动：
x-xxx-xxx-xxxx
电话号码：移动
工作：
y-yyy-yyy-yyyy
电话号码：工作
家庭：
z-zzz-zzz-zzzz
电话号码：家庭
1376
1378
1380
1300

图15

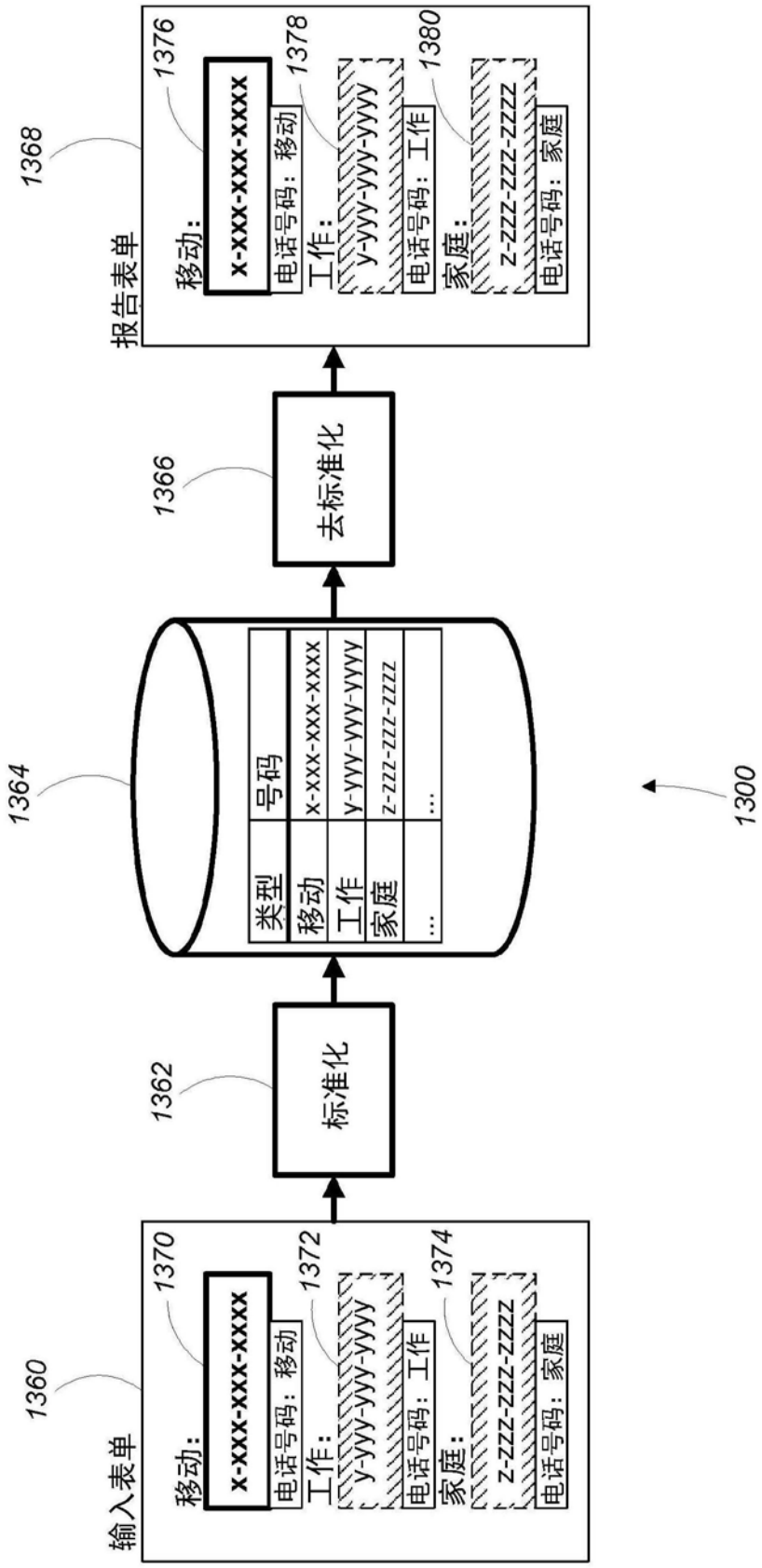

1360
输入表单
移动:
x-xxx-xxx-xxxx
电话号码：移动
工作:
y-yyy-yyy-yyyy
电话号码：工作
家庭:
z-zzz-zzz-zzzz
电话号码：家庭
1370
1372
1374
1362
标准化
1364
类型
号码
移动
x-xxx-xxx-xxxx
工作
y-yyy-yyy-yyyy
家庭
z-zzz-zzz-zzzz
...
1366
去标准化
1368
报告表单
移动:
x-xxx-xxx-xxxx
电话号码：移动
工作:
y-yyy-yyy-yyyy
电话号码：工作
家庭:
z-zzz-zzz-zzzz
电话号码：家庭
1376
1378
1380
1300

图16

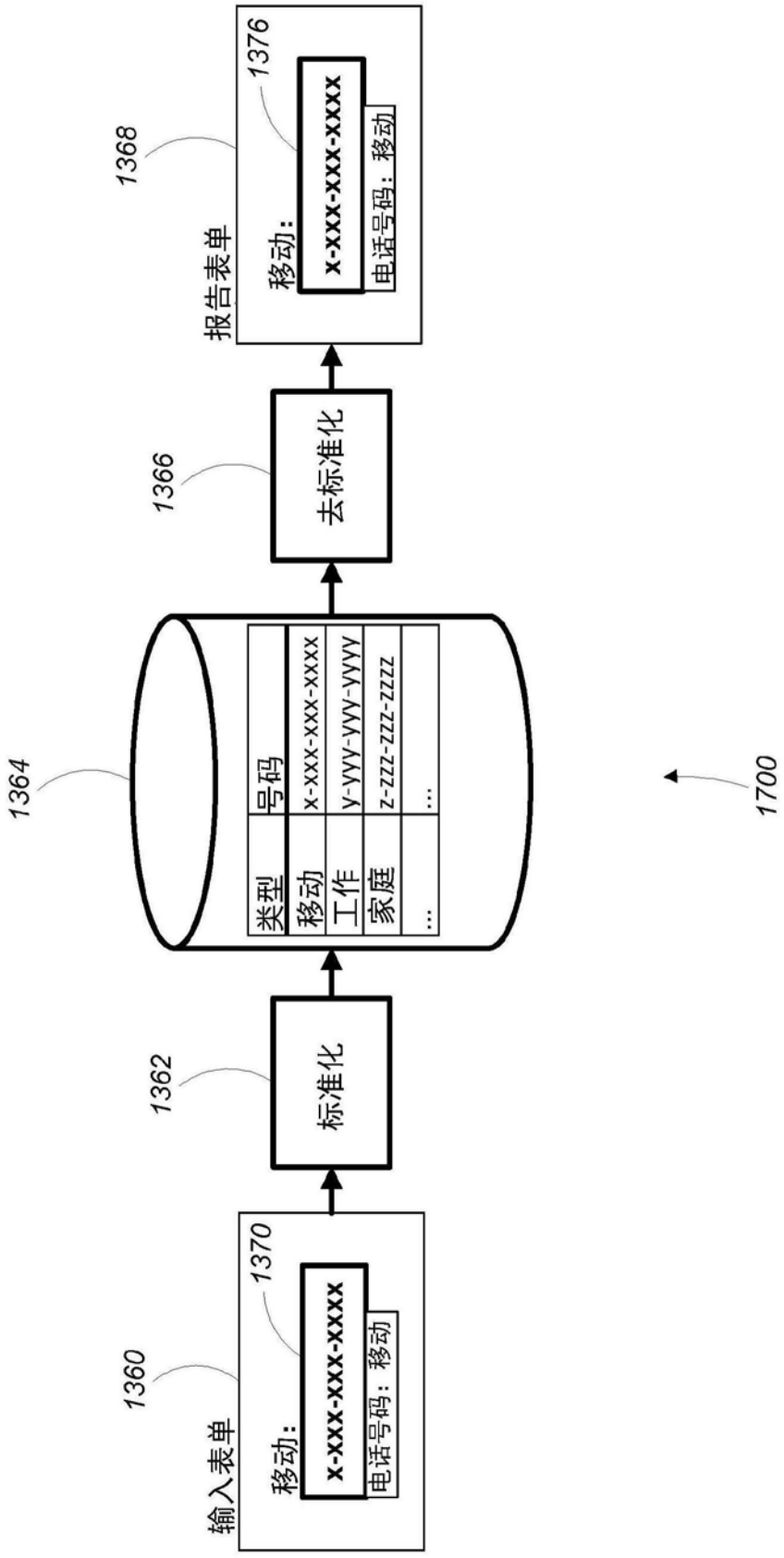

1360
输入表单
移动:
x-xxx-xxx-xxxx
电话号码：移动
1370
1362
标准化
1364
类型
号码
移动
x-xxx-xxx-xxxx
工作
y-yyy-yyy-yyyy
家庭
z-zzz-zzz-zzzz
...
...
1366
去标准化
1368
报告表单
移动:
x-xxx-xxx-xxxx
电话号码：移动
1376
1700

图17