(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 105339940 B
(45)授权公告日 2019.02.26

(21)申请号 201480036405.9

(22)申请日 2014.06.23

(65)同一申请的已公布的文献号
申请公布号 CN 105339940 A

(43)申请公布日 2016.02.17

(30)优先权数据
61/841,045 2013.06.28 US
14/226,557 2014.03.26 US

(85)PCT国际申请进入国家阶段日
2015.12.25

(86)PCT国际申请的申请数据
PCT/US2014/043599 2014.06.23

(87)PCT国际申请的公布数据
WO2014/209848 EN 2014.12.31

(73)专利权人 甲骨文国际公司
地址 美国加利福尼亚

(72)发明人 A·哈蒂 O·迪拉克

(74)专利代理机构 中国国际贸易促进委员会专利商标事务所 11038
代理人 罗亚男

(51)Int.Cl.
G06F 16/22(2019.01)

(56)对比文件
US 2011/0282832 A1,2011.11.17,
US 2011/0282832 A1,2011.11.17,
US 2012/0150791 A1,2012.06.14,
CN 103488697 A,2014.01.01,

审查员 邓丽婉

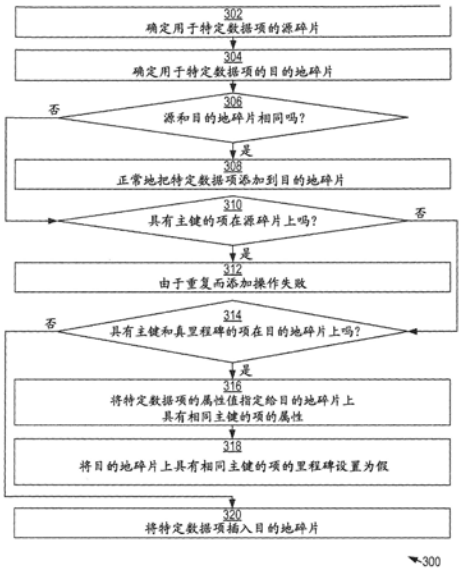
权利要求书10页 说明书19页 附图11页

(54)发明名称

具有碎片的在线添加的朴素的客户端分片

(57)摘要

多个客户端能够相对于碎片系统中的数据项彼此异步地执行操作，而无需那些客户端使用专用锁。其中由于系统中碎片的数量的修改，数据项在一组碎片当中自动重新分布的重新平衡事件可以在客户端不使用专用锁的情况下被执行。甚至当重新平衡过程在系统范围的重新平衡事件期间异步地重新分布至少一些数据项时，客户端也能够继续相对于碎片系统中的至少一些数据项执行操作。能够在不牺牲碎片系统内的数据一致性的情况下获得所有这些好处。



CN 105339940 B

1.一种计算机实现的方法,包括:

确定多碎片系统中碎片的数量已改变;

响应于确定所述多碎片系统中碎片的数量已改变,将客户端从正常状态转变到重新平衡状态,在正常状态中客户端以第一种方式相对于存储在所述多碎片系统中的数据项执行一类型的操作,在重新平衡状态中客户端以与第一种方式不同的第二种方式相对于存储在所述多碎片系统中的所述数据项执行所述类型的操作并且无需获取相对于所述数据项中的任何数据项的专用锁,其中在重新平衡状态中,所述类型的操作基于以下被执行:

所述多碎片系统中的第一碎片,所述第一碎片包括在所述多碎片系统中碎片的数量改变之前存储在所述第一碎片中的第一存储器位置处的一数据项;和

所述多碎片系统中的第二碎片,所述第二碎片包括在所述多碎片系统中碎片的数量改变之后存储在所述第二碎片中的第二存储器位置处的该数据项;

在客户端处于重新平衡状态时,对于一个或多个数据项,确定与在客户端从正常状态转变到重新平衡状态之前所述一个或多个数据项存储在其上的源碎片分开的目的地碎片;以及

在客户端处于重新平衡状态时,将所述一个或多个数据项从源碎片移动到目的地碎片。

2.如权利要求1所述的计算机实现的方法,还包括:

在所述一个或多个数据项从源碎片被移动到目的地碎片时,与所述一个或多个数据项的移动异步地,客户端以第二种方式执行所述类型的操作而无需获取专用锁。

3.如权利要求1或权利要求2所述的计算机实现的方法,还包括:

通过至少部分地基于在碎片数量改变之后多碎片系统中碎片的数量散列数据项的主键,确定用于所述一个或多个数据项的目的地碎片;及

响应于确定所述一个或多个数据项已经从源碎片被移动到目的地碎片,将客户端从重新平衡状态转变到正常状态。

4.如权利要求1所述的计算机实现的方法,还包括:

在所述一个或多个数据项从源碎片被移动到目的地碎片时,客户端至少部分地通过以下操作相对于具有主键的第一数据项以第二种方式执行添加类型的操作:

至少部分地基于在碎片数量改变之前多碎片系统中碎片的数量,确定在碎片数量改变之前第一数据项被存储在其上的源碎片;

至少部分地基于在碎片数量改变之后多碎片系统中碎片的数量,确定在碎片数量改变之后第一数据项要被添加到其上的目的地碎片;

确定源碎片上没有数据项具有第一数据项的主键;

确定目的地碎片上没有数据项既具有真里程碑属性值又具有第一数据项的主键;及

将第一数据项插入目的地碎片。

5.如权利要求1所述的计算机实现的方法,还包括:

在所述一个或多个数据项从源碎片被移动到目的地碎片时,客户端至少部分地通过以下操作相对于具有主键的第一数据项以第二种方式执行添加类型的操作:

至少部分地基于在碎片数量改变之前多碎片系统中碎片的数量,确定在碎片数量改变之前第一数据项被存储在其上的源碎片;

至少部分地基于在碎片数量改变之后多碎片系统中碎片的数量，确定在碎片数量改变之后第一数据项被存储在其上的目的地碎片；

确定源碎片上没有数据项具有第一数据项的主键；

确定目的地碎片上的第二数据项既具有真里程碑属性值又具有第一数据项的主键；

向第二数据项的属性指派第一数据项的属性值；及

将第二数据项的里程碑属性值设定为假。

6.如权利要求1所述的计算机实现的方法，还包括：

在所述一个或多个数据项从源碎片被移动到目的地碎片时，客户端至少部分地通过以下操作相对于具有主键的第一数据项以第二种方式执行更新类型的操作：

至少部分地基于在碎片数量改变之前多碎片系统中碎片的数量，确定在碎片数量改变之前第一数据项被存储在其上的源碎片；

至少部分地基于在碎片数量改变之后多碎片系统中碎片的数量，确定在碎片数量改变之后第一数据项要位于其上的目的地碎片；

确定源碎片上没有数据项具有第一数据项的主键；

确定目的地碎片上的第二数据项既具有假里程碑属性值又具有第一数据项的主键；及

向第二数据项的属性指派第一数据项的属性值。

7.如权利要求1所述的计算机实现的方法，还包括：

在所述一个或多个数据项从源碎片被移动到目的地碎片时，客户端至少部分地通过以下操作相对于具有主键的第一数据项以第二种方式执行更新类型的操作：

至少部分地基于在碎片数量改变之前多碎片系统中碎片的数量，确定在碎片数量改变之前第一数据项被存储在其上的源碎片；

至少部分地基于在碎片数量改变之后多碎片系统中碎片的数量，确定在碎片数量改变之后第一数据项要位于其上的目的地碎片；

确定源碎片上的第二数据项既具有第一数据项的主键又具有假里程碑属性值；

确定目的地碎片上的第三数据项具有第一数据项的主键；

向第三数据项的属性指派第一数据项的属性值；及

将第三数据项的版本号设定为与第二数据项的版本号不同的值。

8.如权利要求1所述的计算机实现的方法，还包括：

在所述一个或多个数据项从源碎片被移动到目的地碎片时，客户端至少部分地通过以下操作相对于具有主键的第一数据项以第二种方式执行更新类型的操作：

至少部分地基于在碎片数量改变之前多碎片系统中碎片的数量，确定在碎片数量改变之前第一数据项被存储在其上的源碎片；

至少部分地基于在碎片数量改变之后多碎片系统中碎片的数量，确定在碎片数量改变之后第一数据项要位于其上的目的地碎片；

确定源碎片上的第二数据项既具有第一数据项的主键又具有假里程碑属性值；

确定目的地碎片上没有数据项具有第一数据项的主键；及

将第一数据项插入目的地碎片，其具有与第二数据项的版本号不同的版本号。

9.如权利要求1所述的计算机实现的方法，还包括：

在所述一个或多个数据项从源碎片被移动到目的地碎片时，客户端至少部分地通过以

下操作相对于具有主键的第一数据项以第二种方式执行删除类型的操作：

至少部分地基于在碎片数量改变之前多碎片系统中碎片的数量，确定在碎片数量改变之前第一数据项被存储在其上的源碎片；

至少部分地基于在碎片数量改变之后多碎片系统中碎片的数量，确定在碎片数量改变之后第一数据项要位于其上的目的地碎片；

确定源碎片上的第二数据项具有第一数据项的主键；

将第一数据项融合到目的地碎片，其具有与第二数据项的版本号不同的版本号；及

将目的地碎片上第一数据项的里程碑属性值设为真。

10.如权利要求1所述的计算机实现的方法，还包括：

在所述一个或多个数据项从源碎片被移动到目的地碎片时，客户端至少部分地通过以下操作相对于具有主键的第一数据项以第二种方式执行删除类型的操作：

至少部分地基于在碎片数量改变之前多碎片系统中碎片的数量，确定在碎片数量改变之前第一数据项被存储在其上的源碎片；

至少部分地基于在碎片数量改变之后多碎片系统中碎片的数量，确定在碎片数量改变之后第一数据项要位于其上的目的地碎片；

确定源碎片上没有数据项具有第一数据项的主键；

确定目的地碎片上的第二数据项具有第一数据项的主键；及

将第二数据项的里程碑属性值设为真。

11.如权利要求1所述的计算机实现的方法，还包括：

在所述一个或多个数据项从源碎片被移动到目的地碎片时，客户端至少部分地通过以下操作相对于主键以第二种方式执行获得类型的操作：

至少部分地基于在碎片数量改变之前多碎片系统中碎片的数量，确定在碎片数量改变之前具有所述主键的数据项被存储在其上的源碎片；

至少部分地基于在碎片数量改变之后多碎片系统中碎片的数量，确定在碎片数量改变之后具有所述主键的数据项要位于其上的目的地碎片；

确定目的地碎片上的第一数据项具有所述主键和假里程碑属性值；及

读取第一数据项的属性值。

12.如权利要求1所述的计算机实现的方法，还包括：

在所述一个或多个数据项从源碎片被移动到目的地碎片时，客户端至少部分地通过以下操作相对于主键以第二种方式执行获得类型的操作：

至少部分地基于在碎片数量改变之前多碎片系统中碎片的数量，确定在碎片数量改变之前具有所述主键的数据项被存储在其上的源碎片；

至少部分地基于在碎片数量改变之后多碎片系统中碎片的数量，确定在碎片数量改变之后具有所述主键的数据项要位于其上的目的地碎片；

确定目的地碎片上没有数据项具有所述主键；

确定源碎片上的第一数据项具有所述主键和假里程碑属性值；及

读取第一数据项的属性值。

13.如权利要求1所述的计算机实现的方法，还包括：

在所述一个或多个数据项从源碎片被移动到目的地碎片时，客户端至少部分地通过以

下操作相对于主键以第二种方式执行获得类型的操作：

至少部分地基于在碎片数量改变之前多碎片系统中碎片的数量，确定在碎片数量改变之前具有所述主键的数据项被存储在其上的源碎片；

至少部分地基于在碎片数量改变之后多碎片系统中碎片的数量，确定在碎片数量改变之后具有所述主键的数据项要位于其上的目的地碎片；

在第一时刻，确定目的地碎片上没有数据项具有所述主键；

确定源碎片上没有数据项具有所述主键；

在第二时刻，确定目的地碎片上的第一数据项具有所述主键和假里程碑属性值；及

读取第一数据项的属性值。

14.如权利要求1所述的计算机实现的方法，其中将所述一个或多个数据项从源碎片移动到目的地碎片包括：

递增系统范围的版本号；及

对于其版本号小于系统范围的版本号的每个第一数据项，至少部分地基于在碎片数量改变之后多碎片系统中碎片的数量，确定在碎片数量改变之后第一数据项要位于其上的第一目的地碎片；

对于具有小于系统范围的版本号的版本号并且位于与在碎片数量改变之后第二数据项要位于其上的第二目的地碎片不同的源碎片上的每个第二数据项，将具有系统范围的版本号的第二数据项的拷贝插入第二目的地碎片；

对于具有小于系统范围的版本号的版本号并且已经位于在碎片数量改变之后第三数据项要位于其上的第三目的地碎片上的每个第三数据项，将第三数据项的版本号改变为系统范围的版本号。

15.如权利要求1所述的计算机实现的方法，其中将所述一个或多个数据项从源碎片移动到目的地碎片包括递增系统范围的版本号，并且还包括：

在所述一个或多个数据项的移动完成时确定针对多碎片系统中的一个或多个碎片的待处理的一组查询；

等待这组查询中的所有查询完成；

在这组查询中的所有查询都完成之后，从多碎片系统中的所有碎片中去除其版本号属性值与系统范围的版本号不同的所有数据项拷贝；及

在这组查询中的所有查询都完成之后，从多碎片系统中的所有碎片中去除具有真里程碑属性值的所有数据项拷贝。

16.如权利要求1所述的计算机实现的方法，还包括：

在所述一个或多个数据项从源碎片被移动到目的地碎片时，并且对于多碎片系统中的每个碎片，利用既位于那个碎片上并且又满足指定的查询标准的数据项为那个碎片填充预备结果队列；

对于每个碎片的预备结果队列，至少部分地基于在那个碎片的预备结果队列中的数据项的主键，在那个碎片的预备结果队列中排序数据项；

在每个碎片的预备结果队列为空之前，重复地执行操作，包括：

从包括当前位于多碎片系统中所有碎片的预备结果队列顶部的数据项的集合选择第一数据项，

将不具有真里程碑属性值的每个第一数据项添加到最终的结果集，及

从每个碎片的预备结果队列中去除其主键与第一数据项的主键匹配的所有数据项拷贝；及

在每个碎片的预备结果队列为空之后，返回最终的结果集中的数据项作为查询结果。

17.如权利要求16所述的计算机实现的方法，其中从包括当前位于多碎片系统中所有碎片的预备结果队列顶部的数据项的集合选择第一数据项包括：

从该集合选择具有该集合中数据项的主键中的最小主键的数据项的子集；及

选择具有该子集中数据项的版本号属性值中的最大版本号属性值的数据项，作为第一数据项。

18.如权利要求1所述的计算机实现的方法，还包括：

在所述一个或多个数据项从源碎片被移动到目的地碎片时，客户端至少部分地通过以下操作尝试相对于具有主键的第一数据项以第二种方式执行删除类型的操作并且失败：

至少部分地基于在碎片数量改变之前多碎片系统中碎片的数量，确定在碎片数量改变之前第一数据项应该已经位于其上的源碎片；

至少部分地基于在碎片数量改变之后多碎片系统中碎片的数量，确定在碎片数量改变之后第一数据项要位于其上的目的地碎片；

确定源碎片上没有数据项具有第一数据项的主键；

确定目的地碎片上没有数据项具有第一数据项的主键；及

由于第一数据项不存在而生成指示删除操作失败的数据。

19.一种计算机系统，包括：

存储数据项的多个数据库碎片；

多个客户端，被配置为响应于所述多个数据库碎片中的碎片的数量改变，从正常状态转变到重新平衡状态，在正常状态中，所述多个客户端当中的每个客户端被配置为相对于数据项以第一种方式执行一类型的操作，在重新平衡状态中，所述多个客户端当中的每个客户端被配置为相对于数据项以与第一种方式不同的第二种方式执行所述类型的操作并且无需获取相对于所述数据项中的任何数据项的专用锁，其中在重新平衡状态中，所述类型的操作基于以下被执行：

所述多个数据库碎片中的第一碎片，所述第一碎片包括在所述多个数据库碎片中的碎片的数量改变之前存储在所述第一碎片中的第一存储器位置处的一数据项；和

所述多个数据库碎片中的第二碎片，所述第二碎片包括在所述多个数据库碎片中的碎片的数量改变之后存储在所述第二碎片中的第二存储器位置处的该数据项；及

至少一个计算设备，被配置为在所述多个客户端处于重新平衡状态时，并且对于一个或多个数据项，确定与在所述多个客户端当中的每个客户端处于正常状态时所述一个或多个数据项存储在其上的源碎片分开的目的地碎片，及在所述多个客户端当中的每个客户端处于重新平衡状态时将所述一个或多个数据项从源碎片移动到目的地碎片。

20.如权利要求19所述的计算机系统，其中：

在所述一个或多个数据项从源碎片被移动到目的地碎片时，与所述一个或多个数据项的移动异步地，客户端以第二种方式执行所述类型的操作而无需获取专用锁。

21.如权利要求19或权利要求20所述的计算机系统，其中，所述至少一个计算设备还被

配置为：

通过至少部分地基于在碎片数量改变之后所述多个数据库碎片中的碎片的数量散列数据项的主键，确定用于所述一个或多个数据项的目的地碎片；及

响应于确定所述一个或多个数据项已经从源碎片被移动到目的地碎片，将客户端从重新平衡状态转变到正常状态。

22.如权利要求19所述的计算机系统，其中：

在所述一个或多个数据项从源碎片被移动到目的地碎片时，客户端至少部分地通过以下操作相对于具有主键的第一数据项以第二种方式执行添加类型的操作：

至少部分地基于在碎片数量改变之前多碎片系统中碎片的数量，确定在碎片数量改变之前第一数据项被存储在其上的源碎片；

至少部分地基于在碎片数量改变之后所述多个数据库碎片中的碎片的数量，确定在碎片数量改变之后第一数据项要被添加到其上的目的地碎片；

确定源碎片上没有数据项具有第一数据项的主键；

确定目的地碎片上没有数据项既具有真里程碑属性值又具有第一数据项的主键；及

将第一数据项插入目的地碎片。

23.如权利要求19所述的计算机系统，其中：

在所述一个或多个数据项从源碎片被移动到目的地碎片时，客户端至少部分地通过以下操作相对于具有主键的第一数据项以第二种方式执行添加类型的操作：

至少部分地基于在碎片数量改变之前所述多个数据库碎片中的碎片的数量，确定在碎片数量改变之前第一数据项被存储在其上的源碎片；

至少部分地基于在碎片数量改变之后所述多个数据库碎片中的碎片的数量，确定在碎片数量改变之后第一数据项要被添加到其上的目的地碎片；

确定源碎片上没有数据项具有第一数据项的主键；

确定目的地碎片上的第二数据项既具有真里程碑属性值又具有第一数据项的主键；

向第二数据项的属性指派第一数据项的属性值；及

将第二数据项的里程碑属性值设定为假。

24.如权利要求19所述的计算机系统，其中：

在所述一个或多个数据项从源碎片被移动到目的地碎片时，客户端至少部分地通过以下操作相对于具有主键的第一数据项以第二种方式执行更新类型的操作：

至少部分地基于在碎片数量改变之前所述多个数据库碎片中的碎片的数量，确定在碎片数量改变之前第一数据项被存储在其上的源碎片；

至少部分地基于在碎片数量改变之后所述多个数据库碎片中的碎片的数量，确定在碎片数量改变之后第一数据项要位于其上的目的地碎片；

确定源碎片上没有数据项具有第一数据项的主键；

确定目的地碎片上的第二数据项既具有假里程碑属性值又具有第一数据项的主键；及

向第二数据项的属性指派第一数据项的属性值。

25.如权利要求19所述的计算机系统，其中：

在所述一个或多个数据项从源碎片被移动到目的地碎片时，客户端至少部分地通过以下操作相对于具有主键的第一数据项以第二种方式执行更新类型的操作：

至少部分地基于在碎片数量改变之前所述多个数据库碎片中的碎片的数量，确定在碎片数量改变之前第一数据项应该已经位于其上的源碎片；

至少部分地基于在碎片数量改变之后所述多个数据库碎片中的碎片的数量，确定在碎片数量改变之后第一数据项要位于其上的目的地碎片；

确定源碎片上的第二数据项既具有第一数据项的主键又具有假里程碑属性值；

确定目的地碎片上的第三数据项具有第一数据项的主键；

向第三数据项的属性指派第一数据项的属性值；及

将第三数据项的版本号设定为与第二数据项的版本号不同的值。

26.如权利要求19所述的计算机系统，其中：

在所述一个或多个数据项从源碎片被移动到目的地碎片时，客户端至少部分地通过以下操作相对于具有主键的第一数据项以第二种方式执行更新类型的操作：

至少部分地基于在碎片数量改变之前所述多个数据库碎片中的碎片的数量，确定在碎片数量改变之前第一数据项被存储在其上的源碎片；

至少部分地基于在碎片数量改变之后所述多个数据库碎片中的碎片的数量，确定在碎片数量改变之后第一数据项要位于其上的目的地碎片；

确定源碎片上的第二数据项既具有第一数据项的主键又具有假里程碑属性值；

确定目的地碎片上没有数据项具有第一数据项的主键；及

将第一数据项插入目的地碎片，其具有与第二数据项的版本号不同的版本号。

27.如权利要求19所述的计算机系统，其中：

在所述一个或多个数据项从源碎片被移动到目的地碎片时，客户端至少部分地通过以下操作相对于具有主键的第一数据项以第二种方式执行删除类型的操作：

至少部分地基于在碎片数量改变之前所述多个数据库碎片中的碎片的数量，确定在碎片数量改变之前第一数据项被存储在其上的源碎片；

至少部分地基于在碎片数量改变之后所述多个数据库碎片中的碎片的数量，确定在碎片数量改变之后第一数据项要位于其上的目的地碎片；

确定源碎片上的第二数据项具有第一数据项的主键；

将第一数据项融合到目的地碎片，其具有与第二数据项的版本号不同的版本号；及

将目的地碎片上第一数据项的里程碑属性值设为真。

28.如权利要求19所述的计算机系统，其中：

在所述一个或多个数据项从源碎片被移动到目的地碎片时，客户端至少部分地通过以下操作相对于具有主键的第一数据项以第二种方式执行删除类型的操作：

至少部分地基于在碎片数量改变之前所述多个数据库碎片中的碎片的数量，确定在碎片数量改变之前第一数据项被存储在其上的源碎片；

至少部分地基于在碎片数量改变之后所述多个数据库碎片中的碎片的数量，确定在碎片数量改变之后第一数据项要位于其上的目的地碎片；

确定源碎片上没有数据项具有第一数据项的主键；

确定目的地碎片上的第二数据项具有第一数据项的主键；及

将第二数据项的里程碑属性值设为真。

29.如权利要求19所述的计算机系统，其中：

在所述一个或多个数据项从源碎片被移动到目的地碎片时，客户端至少部分地通过以下操作相对于主键以第二种方式执行获得类型的操作：

至少部分地基于在碎片数量改变之前所述多个数据库碎片中的碎片的数量，确定在碎片数量改变之前具有所述主键的数据项被存储在其上的源碎片；

至少部分地基于在碎片数量改变之后所述多个数据库碎片中的碎片的数量，确定在碎片数量改变之后具有所述主键的数据项要位于其上的目的地碎片；

确定目的地碎片上的第一数据项具有所述主键和假里程碑属性值；及

读取第一数据项的属性值。

30.如权利要求19所述的计算机系统，其中：

在所述一个或多个数据项从源碎片被移动到目的地碎片时，客户端至少部分地通过以下操作相对于主键以第二种方式执行获得类型的操作：

至少部分地基于在碎片数量改变之前所述多个数据库碎片中的碎片的数量，确定在碎片数量改变之前具有所述主键的数据项被存储在其上的源碎片；

至少部分地基于在碎片数量改变之后所述多个数据库碎片中的碎片的数量，确定在碎片数量改变之后具有所述主键的数据项要位于其上的目的地碎片；

确定目的地碎片上没有数据项具有所述主键；

确定源碎片上的第一数据项具有所述主键和假里程碑属性值；及

读取第一数据项的属性值。

31.如权利要求19所述的计算机系统，其中：

在所述一个或多个数据项从源碎片被移动到目的地碎片时，客户端至少部分地通过以下操作相对于主键以第二种方式执行获得类型的操作：

至少部分地基于在碎片数量改变之前所述多个数据库碎片中的碎片的数量，确定在碎片数量改变之前具有所述主键的数据项被存储在其上的源碎片；

至少部分地基于在碎片数量改变之后所述多个数据库碎片中的碎片的数量，确定在碎片数量改变之后具有所述主键的数据项要位于其上的目的地碎片；

在第一时刻，确定目的地碎片上没有数据项具有所述主键；

确定源碎片上没有数据项具有所述主键；

在第二时刻，确定目的地碎片上的第一数据项具有所述主键和假里程碑属性值；及

读取第一数据项的属性值。

32.如权利要求19所述的计算机系统，其中将所述一个或多个数据项从源碎片移动到目的地碎片包括：

递增系统范围的版本号；及

对于其版本号小于系统范围的版本号的每个第一数据项，至少部分地基于在碎片数量改变之后所述多个数据库碎片中的碎片的数量，确定在碎片数量改变之后第一数据项要位于其上的第一目的地碎片；

对于具有小于系统范围的版本号的版本号并且位于与在碎片数量改变之后第二数据项要位于其上的第二目的地碎片不同的源碎片上的每个第二数据项，将具有系统范围的版本号的第二数据项的拷贝插入第二目的地碎片；

对于具有小于系统范围的版本号的版本号并且已经位于在碎片数量改变之后第三数

据项要位于其上的第三目的地碎片上的每个第三数据项，将第三数据项的版本号改变为系统范围的版本号。

33.如权利要求19所述的计算机系统，其中将所述一个或多个数据项从源碎片移动到目的地碎片包括递增系统范围的版本号，并且其中，所述至少一个计算设备还被配置为：

在所述一个或多个数据项的移动完成时确定针对所述多个数据库碎片中的一个或多个碎片的待处理的一组查询；

等待这组查询中的所有查询完成；

在这组查询中的所有查询都完成之后，从所述多个数据库碎片中的所有碎片中去除其版本号属性值与系统范围的版本号不同的所有数据项拷贝；及

在这组查询中的所有查询都完成之后，从所述多个数据库碎片中的所有碎片中去除具有真里程碑属性值的所有数据项拷贝。

34.如权利要求19所述的计算机系统，其中所述至少一个计算设备还被配置为：

在所述一个或多个数据项从源碎片被移动到目的地碎片时，并且对于所述多个数据库碎片中的每个碎片，利用既位于那个碎片上并且又满足指定的查询标准的数据项为那个碎片填充预备结果队列；

对于每个碎片的预备结果队列，至少部分地基于在那个碎片的预备结果队列中的数据项的主键，在那个碎片的预备结果队列中排序数据项；

在每个碎片的预备结果队列为空之前，重复地执行操作，包括：

从包括当前位于所述多个数据库碎片中的所有碎片的预备结果队列顶部的数据项的集合选择第一数据项，

将不具有真里程碑属性值的每个第一数据项添加到最终的结果集，及

从每个碎片的预备结果队列中去除其主键与第一数据项的主键匹配的所有数据项拷贝；及

在每个碎片的预备结果队列为空之后，返回最终的结果集中的数据项作为查询结果。

35.如权利要求34所述的计算机系统，其中从包括当前位于多个数据库碎片系统中所有碎片的预备结果队列顶部的数据项的集合选择第一数据项包括：

从该集合选择具有该集合中数据项的主键中的最小主键的数据项的子集；及

选择具有该子集中数据项的版本号属性值中的最大版本号属性值的数据项，作为第一数据项。

36.如权利要求19所述的计算机系统，其中：

在所述一个或多个数据项从源碎片被移动到目的地碎片时，客户端至少部分地通过以下操作尝试相对于具有主键的第一数据项以第二种方式执行删除类型的操作并且失败：

至少部分地基于在碎片数量改变之前所述多个数据库碎片中的碎片的数量，确定在碎片数量改变之前第一数据项被存储在其上的源碎片；

至少部分地基于在碎片数量改变之后所述多个数据库碎片中的碎片的数量，确定在碎片数量改变之后第一数据项要位于其上的目的地碎片；

确定源碎片上没有数据项具有第一数据项的主键；

确定目的地碎片上没有数据项具有第一数据项的主键；及

由于第一数据项不存在而生成指示删除操作失败的数据。

37.一种计算机实现的设备，包括用于执行如权利要求1-18中任何一项所述的计算机实现的方法的装置。

具有碎片的在线添加的朴素的客户端分片

[0001] 版权声明

[0002] 本专利文档公开内容的一部分包含受版权保护的素材。版权拥有者不反对任何人对专利文档或专利公开内容按照在专利商标局的专利文件或记录中出现的那样进行传真复制，但是除此之外在任何情况下都保留所有版权。

[0003] 对相关申请的交叉引用；要求优先权

[0004] 本申请依据35U.S.C.§119要求于2013年6月28日提交的标题为“NAÏVE, CLIENT-SIDE SHARDING WITH ONLINE ADDITION OF SHARDS”的美国临时专利申请序列No.61/841,045的优先权，该申请的全部内容通过引用被结合于此；以及要求于2014年3月26日提交的标题为“NAÏVE,CLIENT SIDE SHARDING WITH ONLINE ADDITION OF SHARDS”的美国专利申请序列No.14/226,557的优先权，该申请全部内容通过引用被结合于此。

背景技术

[0005] 在数据库碎片的系统中，诸如文档的记录或行之类的数据项可以在多个被称为碎片的单独、独立的数据库中分布。在这种系统中，数据项通常不是并且不被允许在碎片当中被复制。因此，在这种系统中，特定数据项在任何给定的时间将只位于几个碎片当中一个上。为了让客户端确定几个碎片当中哪个包含特定数据项，客户端可以向散列函数输入该数据项的主键-该主键唯一地标识特定数据项。散列函数基于主键计算该特定数据项当前存储在其上的碎片的身份。例如，散列函数可以用数值主键除以系统中碎片的数量，然后取余数(基本上是模运算)作为包含特定数据项的碎片的标识符。利用这种散列函数首先确定每个数据项将存储在其上的碎片通常使数据项相对均匀地在碎片当中分布。

[0006] 一旦客户端已经识别出特定数据项位于其上的碎片，客户端就可以相对于该数据项执行操作，诸如读取、删除或更新操作。通常，碎片系统将同时服务于众多的客户端，并且这些客户端每个都可以相对于单独的数据项彼此异步地执行操作。有可能多个客户端无意中尝试同时相对于相同的数据项执行操作。如果这种场景被允许不受阻碍地发生，则数据项会变得被破坏，从而使碎片系统的状态不一致。根据一种方法，为了保证多个客户端将不同时相对于相同的数据项执行操作，设法相对于特定数据项执行操作的客户端可以首先被要求获得关于该特定数据项的专用锁。每个数据项可以与单独的锁关联。如果另一个客户端已经持有关于特定数据项的专用锁，则阻止客户端获取该专用锁；在这种情形之下，设法获得该专用锁的客户端必须等待持有锁的客户端释放该专用锁。在客户端持有关于特定数据项的专用锁时，那个客户端可以相对于该特定数据项执行操作。当客户端相对于特定数据项完成执行操作时，该客户端释放关于该特定数据项的专用锁，由此使该特定数据项可以让其它客户端访问。

[0007] 随着存储在碎片系统内的数据的数量增长，系统中现有碎片的容量可能变得不足以包含要在系统中存储的所有数据。在那些情形之下，会期望向系统添加一个或多个新碎片。新碎片的添加可以涉及新硬件计算和存储设备的添加，以包含和管理新数据库。为了尝

试在碎片当中平衡客户端访问负载，以使得没有一个碎片子集由于客户端请求而变得不成比例地过重负担，新碎片向系统的添加可以促成系统所存储的数据项在扩充的碎片组当中的重新分布。重新分布事件或者重新平衡事件可以使之前存储在一个碎片上的数据项重新定位到另一个碎片，有可能但不一定定位到新添加的碎片。根据一种方法，重新平衡过程可以获得关于要被移动的数据项的专用锁。在获得关于数据项的专用锁之后，重新平衡过程可以将那些数据项从旧的碎片移动到已经被修订的散列函数确定为用作那些数据项的目的地的新碎片。在移动数据项之后，重新平衡过程可以释放关于那些数据项的专用锁。

[0008] 只要专用锁还存在，伴随着它们的使用就有一些缺陷。一个缺陷是，当数据项的专用锁被过程持有时，没有其它过程能够访问该数据项。因此，根据以上讨论的利用锁的重新平衡方法，在重新平衡事件进行时，客户端可能很大程度上不能相对于碎片系统执行操作；在重新平衡过程持有关于数据项的专用锁时，没有客户端能够获得该专用锁。甚至当重新平衡事件不进行时，为了防止并发的多个客户端访问的客户端获取、维护和锁的释放所涉及的开销也会显著而且繁重。即使忽略重新平衡事件的影响，碎片系统中锁的使用也会不利地影响系统效率和性能。还有可能更糟的是，碎片系统中非预期的故障会使持有锁的过程(这可以是，例如，客户端或重新平衡过程)冻结或以别的方式退出正常运行。在这种情形之下，不工作的过程会保留关于特定数据项的专用锁直到某个定时器到期，在那个时候不工作的过程可以被终止，并且它所持有的锁被强制释放。包括客户端或重新平衡过程的其它过程，因此在相对于那个特定数据项继续它们的预期任务之前被迫等待定时器到期。尤其是如果要相对于特定数据项执行的操作仅仅是要相对于多个单独数据项执行的严格有序的操作序列当中的一步时，这种强制等待会使整个碎片系统的性能显著降级。由常常需要被执行的操作的次序所强加的依赖性会引起这些类型复杂因素的级联。

附图说明

[0009] 图1是示出根据本发明的实施例的可扩展碎片系统的例子的框图，其中多个客户端可以访问已经在多个数据库碎片当中分布的数据项；

[0010] 图2是示出根据本发明的实施例的其中系统可以在各个时刻存在的各种状态以及那些状态之间的可能转变的状态图；

[0011] 图3是示出根据本发明的实施例的、用于在处于重新平衡状态时执行添加操作的技术的例子的流程图；

[0012] 图4是示出根据本发明的实施例的、用于在处于重新平衡状态时执行更新操作的技术的例子的流程图；

[0013] 图5是示出根据本发明的实施例的、用于在处于重新平衡状态时执行删除操作的技术的例子的流程图；

[0014] 图6A-6B是示出根据本发明的实施例的、用于在处于重新平衡状态时执行获得操作的技术的例子的流程图；

[0015] 图7是示出根据本发明的实施例的、用于在(最初)处于重新平衡状态时执行重新平衡操作的技术的例子的流程图；

[0016] 图8是示出根据本发明的实施例的、用于在处于重新平衡状态时执行查询操作的技术的例子的流程图；

[0017] 图9绘出了用于实现其中一种实施例的分布式系统的简化图。

[0018] 图10是根据本公开内容的实施例的系统环境的部件的简化框图，通过该环境，由实施例系统的部件提供的服务可以作为云服务提供。图11示出了本发明的各种实施例可以在其中实现的计算机系统的例子。

具体实施方式

[0019] 在以下描述中，为了解释，阐述了具体的细节以便提供对本发明的实施例的透彻理解。但是，很显然，没有这些具体细节本发明也可以实践。

[0020] 根据本发明的实施例，可以使多个客户端能够相对于碎片系统中的数据项彼此异步地执行操作，而无需那些客户端使用专用锁。此外，根据本发明的实施例，可以在客户端不使用专用锁的情况下执行重新平衡事件，其中由于系统中碎片的数量的修改(添加或者即将到来的去除)，数据项在一组碎片当中自动重新分布。此外，在本发明的实施例中，即使当重新平衡过程在系统范围的重新平衡事件期间异步地重新分布至少一些数据项时，客户端也可以继续相对于碎片系统中的至少一些数据项执行操作。提供这些特征的编程代码可以专门位于客户端而不是碎片服务器上。可以在不牺牲碎片系统内的数据一致性的情况下获得所有这些好处。

[0021] 示例可扩展碎片系统

[0022] 图1是示出根据本发明的实施例的可扩展碎片系统100的例子的框图，其中多个客户端可以访问已经在多个数据库碎片当中分布的数据项。碎片系统100包括客户端102A-N和碎片104A-N。系统100中客户端和碎片的数量可以变化。碎片104A-N中的每个碎片可以是不需要知道系统100内任何其它碎片的单独且独立的数据库。例如，碎片104A-N中的每个碎片可以包括单独的数据库服务器和关系数据库。虽然数据库在本文作为具体的例子被讨论，但是本发明的实施例可以应用到除数据库之外的各种类型的数据储存库(例如，轻量级目录访问协议(LDAP)目录、平面文件、关联存储器，等等)。客户端102A-N中的每个客户端可以是能够独立于客户端102A-N中的彼此操作的单独的计算系统。例如，客户端102A-N可以是台式计算机、膝上型计算机、移动设备等等。

[0023] 客户端102A-N可以通过网络106与碎片104A-N交互。网络106可以是或者可以包括，局域网(LAN)、广域网(WAN)和/或互联网。经网络106的通信可以通过一套网络通信协议(诸如以太网、传输控制协议/网际协议(TCP/IP)、超文本传输协议(HTTP)、简单对象访问协议(SOAP)、开放式数据库连接(ODBC)等)来实现。客户端102A-N中的每个客户端可以执行利用散列函数的软件程序的单独实例，以便基于特定数据项的主键计算碎片104A-N当中该特定数据项已经在其上存储或者要存储到其上的特定碎片的身份。这种数据项可以是对相似属性集拥有不同值的单独记录。至少在一种实施例中，这种数据项可以作为一个或多个关系数据库表中的单独行被存储在碎片104A-N内。利用以下讨论的在重新平衡事件期间适用于系统100的非常具体的例外，每个数据项在任何特定的时刻仅位于碎片104A-N当中的一个上。重新平衡事件潜在地会引起各个数据项从一个碎片重新定位到另一个碎片。在本发明的实施例中，在重新平衡事件期间客户端的操作执行中所涉及的活动顺序可以与在重新平衡事件之外发生的“正常”系统状态期间客户端的相同类型操作执行中所涉及的活动顺序不同。

[0024] 一般而言，为了相对于已经存储在碎片104A-N中的特定的一个碎片上的数据项执行操作，客户端102A-N中的特定客户端可以首先(例如，基于散列函数)确定数据项当前存储在其上的碎片。在重新平衡事件期间，特定客户端可以执行检查以确保该特定客户端将对数据项的正确拷贝操作，以便对特定客户端的活动补偿数据项可能被异步重新定位或者删除的可能性。虽然在正常系统状态期间，数据项只有一个拷贝可以存在于系统100中任何地方，但是在重新平衡事件期间，可能有数据项的多个拷贝(不同版本)临时存在于系统100内。在执行这些检查时，特定客户端可以使用与每个数据项的每个拷贝一起存储的版本信息(以下更详细地讨论)。至少部分地基于这种版本信息，特定客户端可以确保如果操作被执行的话，操作将相对于存储在数据项最近被置于其上的碎片上的数据项的拷贝来执行。在重新平衡事件期间特定客户端的操作执行可以涉及数据项的新拷贝在与数据项的另一拷贝之前存在于其上的碎片分开的碎片上的创建。在实施例中，该操作可以涉及相对于数据项执行一个或多个指令。例如，这种指令可以采取查询语言指令(结构化查询语言(SQL)指令仅仅是一种具体的可能性)的形式。在特定客户端相对于数据项的最新拷贝执行操作之后，可以执行清除操作以便从数据项不应当在其上存在的碎片中去除数据项的旧的和过时的版本。一般而言，系统100内的数据一致性可以以这种方式来维持，而无需客户端102A-N使用锁。

[0025] 根据本发明的实施例，因为可以将碎片添加到系统100(或从其去除)，所以系统100是可扩展的。碎片104A-N的数量的修改会引起在系统100内发生重新平衡事件。在重新平衡事件期间，重新平衡过程可以基于碎片的新数量而不是旧数量来重新散列每个数据项的主键，由此确定在重新平衡事件结束时那个数据项应当位于其上的碎片(不同的或相同的)的身份。然后，重新平衡过程可以利用以下更详细讨论的技术从碎片到碎片重新定位数据项。重新定位可以涉及在重新平衡事件之外将不在系统100内存在的数据项的拷贝的创建和删除。重新平衡过程可以与在客户端102A-N上执行的软件异步地执行。有利地，至少部分地通过调整重新平衡操作期间客户端102A-N执行的活动的序列，客户端102A-N可以在重新平衡操作期间继续相对于数据项执行操作。

[0026] 版本和里程碑属性

[0027] 如以上所提到的，在本发明的实施例中，特定数据项的每个拷贝可以与版本信息相关联地存储。在实施例中，这种版本信息可以采取系统范围的版本号的形式，该版本号在每次重新平衡事件在系统100中发生时递增。因此，例如，如果现有的数据项的版本号是3，并且如果在数据项的新拷贝在重新平衡事件期间在不同的碎片上创建时系统的当前版本号是4，则那个数据项的新拷贝将具有版本号4。通过检查数据项的两个单独拷贝的版本信息，客户端可以确定哪个拷贝是更新的版本，并且可以相对于那个拷贝执行操作，其中所述两个单独拷贝可能在重新平衡事件期间存在。此外，通过检查数据项的两个单独拷贝的版本信息，重新平衡过程可以确定哪个拷贝是较旧的版本，并且因此应当从那个拷贝位于其上的碎片被去除，其中所述两个单独拷贝可能在重新平衡事件期间存在。

[0028] 客户端执行的操作可以涉及从碎片删除数据项。在实施例中，客户端相对于数据项的删除操作的执行不立即从数据项位于其上的碎片去除该数据项的所有痕迹。相反，在实施例中，每个数据项都具有被称为“里程碑”的属性，其值可以被设定为真(如果在那个碎片上的数据项的拷贝已经被删除)或假(如果在那个碎片上的数据项的拷贝还没有被删

除)。数据项的里程碑属性设定为“真”避免否则的话在数据项在碎片上不存在时可能发生的歧义情况,其中数据项在碎片上不存在或者是由于删除或者是由于作为重新平衡事件的一部分被移动到另一碎片。在实施例中,只有重新平衡过程(而不是客户端)被允许从碎片完全去除数据项。

[0029] 系统状态转变

[0030] 如以上所讨论的,在不同的时刻,系统100可以处于其中它执行重新平衡事件的状态,或者系统100可以处于在这种重新平衡事件之外的状态。而且如以上所讨论的,在重新平衡事件期间客户端102A-N作为相对于数据项的操作的一部分执行的活动序列可以与在重新平衡事件不发生的状态期间客户端102A-N作为相同类型操作的一部分执行的活动序列不同。以简化的方式来看,在系统100不经历重新平衡事件时存在的“正常”状态期间,客户端102A-N可以利用简单、高效的“朴素”技术相对于数据项执行操作,该技术可能缺乏防止某些类型不一致的安全措施。作为对比,在系统100经历重新平衡事件时存在的状态期间,客户端102A-N可以利用强加这种安全措施的更复杂、更谨慎的技术执行相同类型的操作。在重新平衡事件之外可能没必要的这种安全措施在重新平衡事件期间会是适合的。

[0031] 图2是示出根据本发明的实施例的其中客户端102A-N可以在各个时刻存在的各种状态200以及那些状态之间的可能转变的状态图。状态200可以包括正常状态202、进入重新平衡状态204、重新平衡状态206、进入清除状态208、清除状态210和离开重新平衡状态212。在一种实施例中,只有在正常状态202和重新平衡状态206期间允许客户端102A-N相对于存储在碎片104A-N内的数据项执行操作,这些状态有可能是客户端102A-N在大部分时间处于其中的状态。虽然客户端102A-N通常将处于相同的状态,但是如从以下讨论将看到的,那些客户端中的单独客户端可以在状态转变期间暂时处于不同状态。

[0032] 根据实施例,客户端102A-N可以在正常状态202被初始化。客户端102A-N可以在处于正常状态202时开始新操作。通知机制可以通知客户端102A-N中的每个客户端一个或多个碎片已经添加到系统100或者将要从系统100去除。响应于这种通知,客户端102A-N中的每个客户端可以等待其待处理的操作完成,并且然后该客户端可以转变到进入重新平衡状态204。在处于进入重新平衡状态204时,客户端102A-N不开始新操作;客户端102A-N可以将要执行的操作排队。当客户端102A-N中的所有客户端都转变到进入重新平衡状态204时,客户端102A-N中的每个客户端可以转变到重新平衡状态206。一旦客户端102A-N已经进入重新平衡状态206,就可以继续进行重新平衡过程以将数据项从源碎片移动到目的地碎片。在本发明的实施例中,从那个时刻向前新数据项拷贝将变得与之关联的当前版本号可以在客户端102A-N进入重新平衡状态206时递增。在处于重新平衡状态206时,客户端102A-N可以开始新操作(有可能包括排队的操作)。当重新平衡过程已经将要被重新定位的所有数据项移动了时,通知机制可以通知客户端102A-N中的每个客户端这个事实。响应于这种通知,客户端102A-N中的每个客户端可以等待其待处理的操作完成,并且然后那个客户端可以转变到进入清除状态208。在处于进入清除状态208时,客户端102A-N不开始新的操作;客户端102A-N可以将要执行的操作排队。当客户端102A-N中的所有客户端都已经转变到进入清除状态208时,客户端102A-N中的每个客户端都可以转变到清除状态210。一旦客户端102A-N进入清除状态210,重新平衡过程就可以继续进行以从碎片去除不应当再在任何碎片上存在的数据项拷贝。在实施例中,当客户端102A-N处于清除状态210时,重新平衡过程可以从

碎片去除具有“真”里程碑属性值的所有数据项拷贝。此外，在实施例中，当客户端102A-N处于清除状态210时，重新平衡过程可以从碎片去除其版本号属性值小于系统的当前版本号的所有数据项拷贝。在处于清除状态210期间，客户端102A-N不开始新的操作；客户端102A-N可以将要执行的操作排队。当重新平衡过程已经去除要被去除的所有数据项拷贝时，通知机制可以通知客户端102A-N中的每个客户端这个事实。响应于这种通知，客户端102A-N中的每个客户端可以转变到离开重新平衡状态212。在处于离开重新平衡状态212时，客户端102A-N不开始新的操作；客户端102A-N可以将要执行的操作排队。当客户端102A-N中的所有客户端都已经转变到离开重新平衡状态212时，客户端102A-N中每个客户端都可以转变回正常状态202。在处于正常状态202时，客户端102A-N可以开始新的操作(有可能包括排队的操作)。

[0033] 如以上所讨论的，当客户端102A-N处于重新平衡状态206时，客户端102A-N可以以比客户端102A-N在处于正常状态202时执行相同类型操作的方式更谨慎的方式执行操作。不管可以从碎片到碎片重新定位数据项的重新平衡过程的同时异步执行如何，客户端102A-N能够在处于重新平衡状态206时执行操作的方式都可以保证数据一致性-这是在正常状态202下不存在的担心。以下讨论在处于重新平衡状态206时用于以这种更谨慎、保证一致性的方式执行各种不同类型的操作的技术。

[0034] 添加操作

[0035] 图3是示出根据本发明的实施例的、用于在处于重新平衡状态时执行添加操作的技术300的例子的流程图。虽然技术300被说明为包括以特殊次序执行的特殊活动，但是本发明的备选实施例可以涉及具有附加的、更少的或者不同的活动的技术。客户端102A-N中的任何客户端都可以执行技术300。在方框302中，客户端可以确定在碎片数量改变之前特定数据项应该已经位于其上的源碎片的身份。例如可以通过计算该特定数据项的主键对旧的碎片数量取模来实现这种确定。在方框304中，客户端可以确定在碎片数量改变之后该特定数据项要位于其上的目的地碎片的身份。例如可以通过计算该特定数据项的主键对新的碎片数量取模来实现这种确定。在方框306中，客户端可以确定源碎片的身份与目的地碎片的身份是否相同。如果身份相同，则控制传递到方框308。否则，控制传递到方框310。

[0036] 在方框308中，客户端可以以用于执行添加操作的正常、朴素和高效的标准方式将特定数据项添加到目的地碎片。在这个时候，技术300终止。

[0037] 作为替代，在方框310中，客户端可以确定在源碎片上是否已经存在具有该特定数据项的主键的数据项。如果在源碎片上已经存在具有该特定数据项的主键的数据项，则控制传递到方框312。否则，控制传递到方框314。

[0038] 在方框312中，客户端可以得出结论，该特定数据项复制已经在碎片系统中存在的数据项，并且可以制止执行添加操作。客户端可以告知用户添加操作由于重复而被阻止。在这个时候，技术300终止。

[0039] 作为替代，在方框314中，客户端可以确定在目的地碎片上是否已经存在具有(a)该特定数据项的主键和(b)为“真”的里程碑属性值的数据项。如果在目的地碎片上已经存在具有(a)该特定数据项的主键和(b)为“真”的里程碑属性值的数据项，则控制传递到方框316。否则，控制传递到方框320。

[0040] 在方框316中，作为原子动作的一部分，客户端可以向(目的地碎片上)具有该特定

数据项的主键的数据项的属性值指派该特定数据项的属性值。这种指派实质上更新了在目的地碎片上存在的数据项。客户端可以向该数据项的版本号属性指派系统的当前版本号。在方框318中，作为同一原子动作的一部分，客户端可以将(目的地碎片上)具有该特定数据项的主键的数据项的里程碑属性值设为“假”。在实施例中，方框316的属性值指派实现了与由方框318的活动所实现的相同的结果，因为在指派之前该特定数据项的里程碑属性值将已经是“假”。在这个时候，技术300终止。

[0041] 作为替代，在方框320中，客户端可以将该特定数据项插入目的地碎片。客户端可以向该特定数据项的版本号属性指派系统的当前版本号。客户端可以告知用户添加操作成功。在这个时候，技术300终止。在本发明的实施例中，方框314-320的活动是作为单个原子操作被执行的。

[0042] 更新操作

[0043] 图4是示出根据本发明的实施例的、用于在处于重新平衡状态时执行更新操作的技术400的例子的流程图。虽然技术400被说明为包括以特殊次序执行的特殊活动，但是本发明的备选实施例可以涉及具有附加的、更少的或者不同的活动的技术。客户端102A-N中的任何客户端都可以执行技术400。在方框402中，客户端可以确定在碎片数量改变之前特定数据项应该已经位于其上的源碎片的身份。例如可以通过计算该特定数据项的主键对旧的碎片数量取模来实现这种确定。在方框404中，客户端可以确定在碎片数量改变之后该特定数据项要位于其上的目的地碎片的身份。例如可以通过计算该特定数据项的主键对新的碎片数量取模来实现这种确定。在方框406中，客户端可以确定源碎片的身份与目的地碎片的身份是否相同。如果身份相同，则控制传递到方框408。否则，控制传递到方框410。

[0044] 在方框408中，客户端可以以用于执行更新操作的正常、朴素和高效的标准方式更新目的地碎片上的该特定数据项。在这个时候，技术400终止。

[0045] 作为替代，在方框410中，客户端可以确定在源碎片上是否已经存在具有该特定数据项的主键的数据项。如果在源碎片上已经存在具有该特定数据项的主键的数据项，则控制传递到方框416。否则，控制传递到方框422。

[0046] 在方框416中，客户端可以确定在目的地碎片上是否已经存在具有该特定数据项的主键的数据项。如果在目的地碎片上已经存在具有该特定数据项的主键的数据项，则控制传递到方框418。否则，控制传递到方框420。

[0047] 在方框418中，客户端可以向(目的地碎片上)具有该特定数据项的主键的数据项的属性值指派该特定数据项的属性值。这种指派实质上更新在目的地碎片上存在的数据项。客户端可以向该数据项的版本号属性指派系统的当前版本号。客户端可以告知用户更新操作成功。在这个时候，技术400终止。

[0048] 作为替代，在方框420中，客户端可以将该特定数据项插入目的地碎片。客户端可以向该特定数据项的版本号属性指派系统的当前版本号。客户端可以告知用户更新操作成功。在这个时候，技术400终止。

[0049] 作为替代，在方框422中，客户端可以确定在目的地碎片上是否已经存在具有(a)该特定数据项的主键和(b)为“假”的里程碑属性值的数据项。如果在目的地碎片上已经存在具有(a)该特定数据项的主键和(b)为“假”的里程碑属性值的数据项，则控制传递到方框418。否则，控制传递到方框428。

[0050] 作为替代，在方框428中，客户端可以制止执行更新操作。客户端可以告知用户更新操作失败(因为没有数据项要更新)。在这个时候，技术400终止。

[0051] 删除操作

[0052] 图5是示出根据本发明的实施例的、用于在处于重新平衡状态时执行删除操作的技术500的例子的流程图。虽然技术500被说明为包括以特殊次序执行的特殊活动，但是本发明的备选实施例可以涉及具有附加的、更少的或者不同的活动的技术。客户端102A-N中的任何客户端都可以执行技术500。在方框502中，客户端可以确定在碎片数量改变之前特定数据项应该已经位于其上的源碎片的身份。例如可以通过计算该特定数据项的主键对旧的碎片数量取模来实现这种确定。在方框504中，客户端可以确定在碎片数量改变之后该特定数据项要位于其上的目的地碎片的身份。例如可以通过计算该特定数据项的主键对新的碎片数量取模来实现这种确定。在方框506中，客户端可以确定源碎片的身份与目的地碎片的身份是否相同。如果身份相同，则控制传递到方框508。否则，控制传递到方框510。

[0053] 在方框508中，客户端可以以用于执行删除操作的正常、朴素和高效的标准方式删除目的地碎片上的特定数据项。在这个时候，技术500终止。

[0054] 作为替代，在方框510中，作为原子动作的一部分，客户端可以确定在源碎片上是否已经存在具有该特定数据项的主键的数据项。如果在源碎片上已经存在具有该特定数据项的主键的数据项，则控制传递到方框512。否则，控制传递到方框516。

[0055] 在方框512中，作为原子动作的一部分，客户端可以将特定数据项融合(upsert)到目的地碎片中。融合被定义为：(1)如果由该主键识别出的对象在数据库(即，目的地碎片)中不存在，则插入，或者(2)如果由该主键识别出的对象在数据库(即，目的地碎片)中存在，则更新。在方框514中，作为原子动作的一部分，客户端可以将该特定数据项的里程碑属性的值设为“真”。客户端可以向该特定数据项的版本号属性指派系统的当前版本号。客户端可以告知用户删除操作成功。在这个时候，技术500终止。

[0056] 作为替代，在方框516中，客户端可以确定在目的地碎片上是否已经存在具有该特定数据项的主键的数据项。如果在目的地碎片上已经存在具有该特定数据项的主键的数据项，则控制传递到方框518。否则，控制传递到方框520。

[0057] 在方框518中，客户端可以(在目的地碎片上)将该数据项的里程碑属性的值设为“真”。客户端可以向该数据项的版本号属性指派系统的当前版本号。客户端可以告知用户删除操作成功。在这个时候，技术500终止。

[0058] 作为替代，在方框520中，客户端可以制止执行删除操作。客户端可以告知用户删除操作失败(因为没有数据项要删除)。在这个时候，技术500终止。

[0059] 获得操作

[0060] 图6A-6B是示出根据本发明的实施例的、用于在处于重新平衡状态时执行获得操作的技术600的例子的流程图。在实施例中，获得操作可以读取并返回具有客户端指派的主键的数据项的属性值。虽然技术600被说明为包括以特殊次序执行的特殊活动，但是本发明的备选实施例可以涉及具有附加的、更少的或者不同的活动的技术。客户端102A-N中的任何客户端都可以执行技术600。首先参照图6A，在方框602中，客户端可以确定在碎片数量改变之前具有指定主键的特定数据项已经位于其上的源碎片的身份。例如可以通过计算该指定主键对旧的碎片数量取模来实现这种确定。在方框604中，客户端可以确定在碎片数量改

变之后具有指定主键的特定数据项要位于其上的目的地碎片的身份。例如可以通过计算该指定主键对新的碎片数量取模来实现这种确定。在方框606中，客户端可以确定源碎片的身份与目的地碎片的身份是否相同。如果身份相同，则控制传递到方框608。否则，控制传递到方框610。

[0061] 在方框608中，客户端可以以用于执行获得操作的正常、朴素和高效的标准方式相对于目的地碎片上具有指定主键的特定数据项执行获得操作。在这个时候，技术600终止。

[0062] 作为替代，在方框610中，客户端可以确定在目的地碎片上是否已经存在具有指定主键的数据项。如果在目的地碎片上已经存在具有指定主键的数据项，则控制传递到方框612。否则，控制传递到方框618。

[0063] 在方框612中，客户端可以确定在目的地碎片上具有指定主键的数据项是否具有“真”里程碑属性值。如果在目的地碎片上具有指定主键的数据项具有“真”里程碑属性值，则控制传递到方框614。否则，控制传递到方框616。

[0064] 在方框614中，客户端可以制止执行获得操作。客户端可以告知用户获得操作失败(因为没有找到具有指定主键的数据项)。在这个时候，技术600终止。

[0065] 作为替代，在方框616中，客户端可以在目的地碎片上读取具有指定主键的数据的属性值。客户端可以将这些属性值呈现给用户，同时告知获得操作成功。在这个时候，技术600终止。

[0066] 作为替代，在方框618中，客户端可以确定在源碎片上是否已经存在具有指定主键的数据项。如果在源碎片上已经存在具有指定主键的数据项，则控制传递到方框620。否则，控制传递到图6B上的方框626。

[0067] 作为替代，在方框620中，客户端可以确定在源碎片上具有指定主键的数据项是否具有“真”里程碑属性值。如果在源碎片上具有指定主键的数据项具有“真”里程碑属性值，则控制传递到方框622。否则，控制传递到方框624。

[0068] 在方框622中，客户端可以制止执行获得操作。客户端可以告知用户获得操作失败(因为没有找到具有指定主键的数据项)。在这个时候，技术600终止。

[0069] 作为替代，在方框624中，客户端可以读取源碎片上具有指定主键的数据的属性值。客户端可以将这些属性值呈现给用户，同时告知获得操作成功。在这个时候，技术600终止。

[0070] 现在参照图6B，作为替代，在方框626中，客户端可以(再次)确定在目的地碎片上现在是否存在具有指定主键的数据项。如果在目的地碎片上现在存在具有指定主键的数据项(例如，作为最近将那个数据项移动到目的地碎片的重新平衡过程的结果)，则控制传递到方框628。否则，控制传递到方框630。

[0071] 在方框628中，客户端可以确定在目的地碎片上具有指定主键的数据项是否具有“真”里程碑属性值。如果在目的地碎片上具有指定主键的数据项具有“真”里程碑属性值，则控制传递到方框630。否则，控制传递到方框632。

[0072] 在方框630中，客户端可以制止执行获得操作。客户端可以告知用户获得操作失败(因为没有找到具有指定主键的数据项)。在这个时候，技术600终止。

[0073] 作为替代，在方框632中，客户端可以读取目的地碎片上具有指定主键的数据的属性值。客户端可以将这些属性值呈现给用户，同时告知获得操作成功。在这个时候，技术600

终止。

[0074] 重新平衡操作

[0075] 图7是示出根据本发明的实施例的、用于在(最初)客户端102A-N处于重新平衡状态时执行重新平衡操作的技术700的例子的流程图。虽然技术700被说明为包括以特殊次序执行的特殊活动,但是本发明的备选实施例可以涉及具有附加的、更少的或者不同的活动的技术。重新平衡过程可以与由客户端102A-N执行其它类型操作异步地执行技术700。

[0076] 在方框702中,系统的当前版本号递增。在方框704中,重新平衡过程可以确定是否有任何碎片仍然包含其版本号属性值小于系统的当前版本号的数据项。如果碎片包含其版本号属性值小于系统的当前版本号的特定数据项,则控制传递到方框706。否则,重新平衡过程可以完成对于这个特定重新平衡事件的重新定位数据项,并且控制传递到方框714。

[0077] 在方框706中,重新平衡过程可以确定在碎片数量改变之后特定数据项要位于其上的目的地碎片的身份。例如可以通过计算该特定数据项的主键对新的碎片数量取模来实现这种确定。在方框708中,重新平衡过程可以确定特定数据项当前位于其上的源碎片的身份与目的地碎片的身份是否相同。如果身份相同,则控制传递到方框710。否则,控制传递到方框712。

[0078] 在方框710中,重新平衡过程可以将系统的当前版本号指派给该特定数据项的版本号属性。在这种情形之下,该特定数据项不需要被重新定位到不同的碎片。控制传递到方框704。

[0079] 作为替代,在方框712中,重新平衡过程可以将该特定数据项插入目的地碎片。在一种实施例中,潜在的冲突可以被忽略。重新平衡过程可以将系统的当前版本号指派给特定数据项的版本号属性。值得注意的是,在实施例中,该特定数据项的旧拷贝可以在源碎片上一直保持到方框716和718的活动被执行。控制传递回方框704。

[0080] 作为替代,在方框714中,重新平衡过程可以确定是否有在重新平衡过程完成重新定位数据项时已经执行的任何查询(例如,来自客户端102A-N的)当前对照任何碎片(例如,碎片104A-N)执行。如果在重新平衡过程完成重新定位数据项时已经执行的至少一个查询当前对照至少一个碎片执行,则控制传递到方框714。否则,控制传递到方框716。

[0081] 在数据项重新分配结束时待处理的所有查询都完成之后,在方框716中,重新平衡过程可以从所有碎片中去除其版本号属性值小于系统的当前版本号的所有数据项拷贝。在方框718中,重新平衡过程可以从所有碎片中去除其里程碑属性值为“真”的所有数据项拷贝。在实施例中,方框716和718的活动可以在以上联系图2讨论的清除阶段210期间被执行。

[0082] 查询操作

[0083] 图8是示出根据本发明的实施例的、用于在处于重新平衡状态时执行查询操作的技术800的例子的流程图。虽然技术800被说明为包括以特殊次序执行的特殊活动,但是本发明的备选实施例可以涉及具有附加的、更少的或者不同的活动的技术。例如,客户端102A-N中的任何客户端都可以执行技术800。技术800可以与技术700并发地执行。在方框802中,对于碎片104A-N当中的每个特定碎片,满足查询指定的过滤标准的该特定碎片的所有数据项都可以被放在用于那个特定碎片的单独的预备结果队列中。每个这种预备结果队列开始可以为空。因此,可以为碎片104A-N中的每个碎片填充单独的预备结果队列。在方框804中,对于碎片104A-N当中每个特定碎片,包含在那个特定碎片的预备结果队列内的数据

项可以至少部分地基于那些数据项的主键来排序。因此，例如，每个预备结果队列可以包含被排序的数据项，以使得那个预备结果队列中具有最小数据项主键的数据项可以在那个预备结果队列的前面。在方框806中，确定是否所有碎片的预备结果队列都为空。如果所有碎片的预备结果队列都为空，则控制传递到方框818。否则，控制传递到方框808。

[0084] 在方框808中，从当前位于每个碎片的预备结果队列顶部的数据项集合，可以选择那个集合中的数据项当中具有最小主键的一个或多个数据项的子集。在方框810中，从在方框808中选定的一个或多个数据项的子集，可以选择具有最大版本号属性值的特定数据项。在方框812中，确定该特定数据项的里程碑属性值是否为“假”。如果该特定数据项的里程碑属性值为“假”，则控制传递到方框814。否则，控制传递到方框816。

[0085] 在方框814中，该特定数据项可以被添加到最终的结果集。最终的结果集开始可以为空。控制传递到方框816。在方框816中，可以从所有碎片的预备结果队列中去除具有与特定数据项相同主键的所有数据项拷贝(包括该特定数据项本身)。这种去除潜在地会使其它数据项上升至那些队列当中一个或多个的顶部。控制传递回方框806。

[0086] 作为替代，在方框818中，最终的结果集中的数据项可以作为查询操作的最终结果被返回。

[0087] 硬件概述

[0088] 图9绘出了用于实现其中一种实施例的分布式系统900的简化图。在所示出的实施例中，分布式系统900包括一个或多个客户端计算设备902、904、906和908，它们被配置为经一个或多个网络910执行和操作诸如web浏览器、专有客户端(例如，Oracle Forms)等等的客户端应用。服务器912可以经由网络910与远程客户端计算设备902、904、906和908通信耦合。

[0089] 在各种实施例中，服务器912可以适于运行由系统的一个或多个部件提供的一个或多个服务或软件应用。在一些实施例中，可以作为基于web的或云服务或者根据软件即服务(SaaS)模型向客户端计算设备902、904、906和/或908的用户提供这些服务。操作客户端计算设备902、904、906和/或908的用户又可以利用一个或多个客户端应用与服务器912交互，以利用由这些部件提供的服务。

[0090] 在图中所绘出的配置中，系统900的软件部件918、920和922被示为在服务器912上实现。在其它实施例中，系统900的一个或多个部件和/或由这些部件提供的服务也可以由客户端计算设备902、904、906和/或908当中的一个或多个实现。然后，操作客户端计算设备的用户可以利用一个或多个客户端应用来使用由这些部件提供的服务。这些部件可以在硬件、固件、软件或其组合中实现。应当认识到，可能有各种不同的系统配置，这些配置可以与分布式系统900不同。因此，图中所示的实施例是用于实现实施例系统的分布式系统的一个例子并且不是要作为限制。

[0091] 客户端计算设备902、904、906和/或908可以是便携手持式设备(例如，iPhone®、蜂窝电话、iPad®、计算平板、个人数字助理(PDA))或可穿戴设备(例如，Google Glass®头戴式显示器)，运行诸如Microsoft Windows Mobile®的软件，和/或诸如iOS、Windows Phone、Android、BlackBerry 10、Palm OS等的各种移动操作系统并且启用互联网、电子邮件、短消息服务(SMS)、BlackBerry®或其它通信协议。客户端计算设

备可以是通用个人计算机，作为例子，包括运行各种版本的Microsoft**Windows®**、Apple **Macintosh®**和/或Linux操作系统的个人计算机和/或膝上型计算机。客户端计算设备可以是运行任何各种商用**UNIX®**或类-UNIX操作系统的工作站计算机，包括但不限于各种GNU/Linux操作系统，诸如Google Chrome OS。作为替代，或者附加地，客户端计算设备902、904、906和908可以是能够经(一个或多个)网络910进行通信的任何其它电子设备，诸如瘦客户端计算机、启用互联网的游戏系统(例如，具有或不具有**Kinect®**手势输入设备的微软Xbox游戏控制台)，和/或个人消息通讯设备。

[0092] 虽然示例性分布式系统900被示为具有四个客户端计算设备，但是其可以支持任何数量的客户端计算设备。诸如带传感器的设备等的其它设备可以与服务器912交互。

[0093] 分布式系统900中的(一个或多个)网络910可以是本领域技术人员熟悉的任何类型的网络，这些网络可以利用各种商用协议当中任何一种支持数据通信，包括但不限于TCP/IP(传输控制协议/网际协议)、SNA(系统网络体系架构)、IPX(互联网分组交换)、AppleTalk等。仅仅作为例子，(一个或多个)网络910可以是局域网(LAN)，诸如基于以太网、令牌环等的网络。(一个或多个)网络910可以是广域网和互联网。它可以包括虚拟网络，包括但不限于虚拟专用网(VPN)、内联网、外联网、公共交换电话网(PSTN)、红外线网络、无线网络(例如，按照电气和电子学研究所(IEEE)902.11协议套件、**Bluetooth®**和/或任何其它无线协议当中任何一个操作的网络)；和/或这些和/或其它网络的任意组合。

[0094] 服务器912可以包括一个或多个通用计算机、专用服务器计算机(作为例子，包括PC(个人计算机)服务器、**UNIX®**服务器、中档服务器、大型计算机、机架式服务器等)，服务器群、服务器群集，或者任何其它适当的布置和/或组合。在各种实施例中，服务器912可以适于运行在以上公开内容中描述的一个或多个服务或软件应用。例如，服务器912可以与用于执行以上根据本公开内容的实施例描述的处理的服务器对应。

[0095] 服务器912可以运行操作系统，包括以上讨论的任何操作系统，以及任何商用的服务器操作系统。服务器912还可以运行任何各种附加的服务器应用和/或中间层应用，包括HTTP(超文本传输协议)服务器、FTP(文件传输协议)服务器、CGI(公共网关接口)服务器、**Java®**服务器、数据库服务器等。示例性数据库服务器包括但不限于可从Oracle，Microsoft、Sybase、IBM(国际商业机器)等商购的那些数据库服务器。

[0096] 在一些实现中，服务器912可以包括一个或多个应用，以分析和整合从客户端计算设备902、904、906和908的用户接收到的数据馈送和/或事件更新。作为例子，数据馈送和/或事件更新可以包括但不限于，**Twitter®**馈送、**Facebook®**更新或者从一个或多个第三方信息源接收的实时更新以及连续数据流，这可以包括与传感器数据应用、金融股票、网络性能测量工具(例如，网络监视和流量管理应用)、点击流分析工具、汽车流量监视等相关的实时事件。服务器912还可以包括一个或多个应用，以经由客户端计算设备902、904、906和908的一个或多个显示设备显示数据馈送和/或实时事件。

[0097] 分布式系统900还可以包括一个或多个数据库914和916。数据库914和916可以驻留在各种位置。作为例子，数据库914和916中的一个或多个可以驻留在服务器912本地的非暂态存储介质上(和/或驻留在服务器912中)。作为替代，数据库914和916可以远离服务器912，并且经基于网络的或专用的连接与服务器912通信。在一组实施例中，数据库914和916可以驻留在存储区域网络(SAN)中。类似地，用于执行服务器912所具有的功能的任何必要

的文件都可以适当地本地存储在服务器912上和/或远程存储。在一组实施例中，数据库914和916可以包括适于响应于SQL格式的命令而存储、更新和检索数据的关系数据库，诸如由Oracle提供的数据库。图10是根据本公开内容的实施例的系统环境1000的一个或多个部件的简化框图，通过该环境，由实施例系统的一个或多个部件提供的服务可以提供作为云服务。在所示出的实施例中，系统环境1000包括可以被用户用来与提供云服务的云基础设施系统1002交互的一个或多个客户端计算设备1004、1006和1008。客户端计算设备可以被配置为操作客户端应用(诸如web浏览器、专有客户端(例如，Oracle Forms)或者某个其它应用)，该客户端应用可以被客户端计算设备的用户用来与云基础设施系统1002交互，以使用由云基础设施系统1002提供的服务。

[0098] 应当认识到，在图中所绘出的云基础设施系统1002可以具有除所绘出的那些之外的其它部件。另外，图中所示的实施例仅仅是可以结合本发明的实施例的云基础设施系统的一个例子。在一些其它实施例中，云基础设施系统1002可以具有比图中所示更多或更少的部件、可以结合两个或更多个部件，或者可以具有部件的不同配置或布置。

[0099] 客户端计算设备1004、1006和1008可以是与以上对902、904、906和908所描述的相似的设备。

[0100] 虽然示例性系统环境1000被示为具有三个客户端计算设备，但是其可以支持任何数量的客户端计算设备。诸如带传感器的设备等的其它设备可以与云基础设施系统1002交互。

[0101] (一个或多个)网络1010可以方便客户端1004、1006和1008与云基础设施系统1002之间的通信和数据交换。每个网络可以是本领域技术人员熟悉的、可以利用各种商用协议中任何一种(包括以上对(一个或多个)网络910所描述的那些)支持数据通信的任何类型的网络。

[0102] 云基础设施系统1002可以包括一个或多个计算机和/或服务器，这可以包括以上对服务器912所描述的那些。

[0103] 在某些实施例中，由云基础设施系统提供的服务可以包括可以让云基础设施系统的用户按需获得的一套服务，诸如在线数据存储和备份解决方案、基于web的电子邮件服务、被托管的办公室套件和文档协作服务、数据库处理、被管理的技术支持服务，等等。由云基础设施系统提供的服务可以动态扩展，以满足其用户的需求。由云基础设施系统提供的服务的具体实例化在本文中被称为“服务实例”。一般而言，经由通信网络(诸如互联网)从云服务提供商的系统对用户可用的任何服务被称为“云服务”。通常，在公共云环境中，构成云服务提供商的系统的服务器和系统与消费者自己的本地服务器和系统不同。例如，云服务提供商的系统可以托管应用，并且用户可以经由诸如互联网的通信网络按需订购并使用该应用。

[0104] 在一些例子中，计算机网络云基础设施中的服务可以包括对由云供应商提供给用户的储存器、被托管的数据库、被托管的web服务器、软件应用或其它服务的受保护的计算机网络访问，或者如以别方式在本领域中已知的。例如，服务可以包括通过互联网对云上的远程储存器的受密码保护的访问。作为另一个例子，服务可以包括被联网的开发人员私有使用的、基于web服务的被托管的关系数据库和脚本语言中间件引擎。作为另一个例子，服务可以包括对在云供应商的web站点上托管的电子邮件软件应用的访问。

[0105] 在某些实施例中，云基础设施系统1002可以包括以自助服务、基于订阅、弹性可扩展、可靠、高度可用并且安全的方式交付给消费者的一套应用、中间件和数据服务供应。这种云基础设施系统的例子是由本受让人提供的Oracle公共云(Oracle Public Cloud)。

[0106] 在各种实施例中，云基础设施系统1002可以适于自动供应、管理和跟踪消费者对由云基础设施系统1002提供的服务的订阅。云基础设施系统1002可以经由不同的部署模型提供云服务。例如，可以根据公共云模型提供服务，其中云基础设施系统1002被出售云服务的组织所有(例如，被Oracle所有)并且使得一般的公共或不同工业企业可以获得该服务。作为另一个例子，可以根据私有云模型提供服务，其中云基础设施系统1002仅仅为单个组织操作并且可以为该组织内的一个或多个实体提供服务。还可以根据社区云模型提供云服务，其中云基础设施系统1002和由云基础设施系统1002提供的服务在相关的社区中被几个组织共享。还可以根据混合云模型提供云服务，这是两个或更多个不同模型的组合。

[0107] 在一些实施例中，由云基础设施系统1002提供的服务可以包括根据软件即服务(SaaS)类别、平台即服务(PaaS)类别、基础设施即服务(IaaS)类别或者包括混合服务的其它服务类别提供的一个或多个服务。经由订阅订单，消费者可以订购由云基础设施系统1002提供的一个或多个服务。然后，云基础设施系统1002执行处理，以提供消费者的订阅订单中的服务。

[0108] 在一些实施例中，由云基础设施系统1002提供的服务可以包括但不限于，应用服务、平台服务和基础设施服务。在一些例子中，可以经由SaaS平台由云基础设施系统提供应用服务。SaaS平台可以被配置为提供属于SaaS类别之下的云服务。例如，SaaS平台可以提供在集成开发和部署平台上建立和交付一套按需应用的能力。SaaS平台可以管理和控制用于提供SaaS服务的底层软件和基础设施。通过利用由SaaS平台提供的服务，消费者可以利用在云基础设施系统上执行的应用。消费者可以获取应用服务，而无需消费者购买单独的许可和支持。各种不同的SaaS服务可以被提供。例子包括但不限于，为大型组织提供用于销量性能管理、企业集成和业务灵活性的解决方案的服务。

[0109] 在一些实施例中，平台服务可以由云基础设施系统经由PaaS平台提供。PaaS平台可以被配置为提供属于PaaS类别之下的云服务。平台服务的例子可以包括但不限于，使组织(诸如Oracle)能够在共享的公共体系架构上整合现有应用的服务，以及建立充分利用由平台提供的共享服务的新应用的能力。PaaS平台可以管理和控制用于提供PaaS服务的底层软件和基础设施。消费者可以获取由云基础设施系统提供的PaaS服务，而无需消费者购买单独的许可和支持。平台服务的例子包括但不限于，Oracle Java云服务(JCS)、Oracle数据库云服务(DBCS)及其它。

[0110] 通过利用由PaaS平台提供的服务，消费者可以采用由云基础设施系统支持的编程语言和工具并且还控制所部署的服务。在一些实施例中，由云基础设施系统提供的平台服务可以包括数据库云服务、中间件云服务(例如，Oracle Fusion Middleware服务)和Java云服务。在一种实施例中，数据库云服务可以支持使组织能够汇聚数据库资源并且以数据库云的形式向消费者提供数据库即服务的共享服务部署模型。中间件云服务可以为消费者提供平台来开发和部署各种商业应用，并且Java云服务可以为消费者提供平台来在云基础设施系统中部署Jave应用。

[0111] 可以由云基础设施系统中的IaaS平台提供各种不同的基础设施服务。基础设施服

务方便对诸如储存器、网络和其它基础计算资源之类的底层计算资源的管理和控制，以便让消费者利用由SaaS平台和PaaS平台提供的服务。

[0112] 在某些实施例中，云基础设施系统1002还可以包括基础设施资源1030，用于提供被用来将各种服务提供给云基础设施系统的消费者的资源。在一种实施例中，基础设施资源1030可以包括诸如服务器、储存器和联网资源之类的硬件的预先集成和优化组合，以便执行由PaaS平台和SaaS平台提供的服务。

[0113] 在一些实施例中，云基础设施系统1002中的资源可以被多个用户共享并且按需动态重新分配。此外，资源可以被分配给不同时区中的用户。例如，云基础设施系统1000可以使第一时区中的第一组用户使用云基础设施系统的资源指定的小时数，然后使相同资源被重新分配给位于不同时区中的另一组用户，由此最大化资源的使用率。

[0114] 在某些实施例中，可以提供多个内部共享服务1032，这些服务被云基础设施系统1002的不同部件或模块或者被由云基础设施系统1002提供的服务共享。这些内部共享服务可以包括但不限于，安全和身份服务、集成服务、企业储存库服务、企业管理器服务、病毒扫描和白名单服务、高可用性、备份和恢复服务、用于启用云支持的服务、电子邮件服务、通知服务、文件传输服务等等。

[0115] 在某些实施例中，云基础设施系统1002可以提供在云基础设施系统中云服务(例如，SaaS、PaaS和IaaS服务)的综合管理。在一种实施例中，云管理功能可以包括用于供应、管理和跟踪由云基础设施系统1002接收的消费者的订阅等等的能力。

[0116] 在一种实施例中，如图中所绘出的，云管理功能可以由一个或多个模块提供，诸如订单管理模块1020、订单编排模块1022、订单供应模块1024、订单管理和监视模块1026，以及身份管理模块1028。这些模块可以包括一个或多个计算机和/或服务器，或者利用一个或多个计算机和/或服务器来提供这些模块，其中计算机和服务器可以是通用计算机、专门服务器计算机、服务器群、服务器群集，或者任何其它适当的布置和/或组合。

[0117] 在示例性操作1034中，使用诸如客户端设备1004、1006或1008之类的客户端设备的消费者可以通过请求由云基础设施系统1002提供的一个或多个服务并且对于由云基础设施系统1002提供的一个或多个服务的订阅下订单来与云基础设施系统1002交互。在某些实施例中，消费者可以访问云用户界面(UI)，云UI 1012、云UI 1014和/或云UI 1016，并且经由这些UI下订阅订单。响应于消费者下订单而由云基础设施系统1002接收的订单信息可以包括识别消费者和消费者想订阅的由云基础设施系统1002提供的一个或多个服务的信息。

[0118] 在消费者下订单之后，订单信息经由云UI 1012、1014和/或1016被接收。

[0119] 在操作1036，订单被存储在订单数据库1018中。订单数据库1018可以是由云基础设施系统1002操作并且与其它系统元件一起操作的几个数据库之一。

[0120] 在操作1038，订单信息被转发到订单管理模块1020。在一些情况中，订单管理模块1020可以被配置为执行与订单相关的计费和记帐功能，诸如验证订单，并且在通过验证时预约订单。

[0121] 在操作1040，关于订单的信息被传送到订单编排模块1022。订单编排模块1022可以利用订单信息来编排用于消费者所下订单的服务和资源的供应。在一些情况中，订单编排模块1022可以编排资源的供应，利用订单供应模块1024的服务支持所订阅的服务。

[0122] 在某些实施例中，订单编排模块1022使得能够管理与每个订单相关联的业务处理并且应用业务逻辑，以确定订单是否应该继续供应。在操作1042，在接收到对新订阅的订单时，订单编排模块1022向订单供应模块1024发送分配并配置履行订阅订单所需的那些资源的请求。订单供应模块1024使得能够分配用于由消费者订购的服务的资源。订单供应模块1024在由云基础设施系统1000提供的云服务与物理实现层之间提供抽象层次，该层抽象被用来供应用于提供所请求服务的资源。因此，订单编排模块1022可以与实现细节隔离，诸如服务和资源是否实际上运行供应或者预先供应并且只在请求时被分配/指派。

[0123] 在操作1044，一旦服务和资源被供应，所提供的服务的通知就可以通过云基础设施系统1002的订单供应模块1024被发送到客户端设备1004、1006和/或1008上的消费者。

[0124] 在操作1046，消费者的订阅订单可以由订单管理和监视模块1026管理和跟踪。在一些情况中，订单管理和监视模块1026可以被配置为收集订阅订单中的服务的使用统计，诸如所使用的储存器的量、所传送的数据的量、用户的数量，以及系统工作时间和系统停机时间的量。

[0125] 在某些实施例中，云基础设施系统1002可以包括身份管理模块1028。身份管理模块1028可以被配置为提供身份服务，诸如在云基础设施系统1002中的访问管理和授权服务。在一些实施例中，身份管理模块1028可以控制关于希望利用由云基础设施系统1002提供的服务的消费者的信息。这种信息可以包括认证这种消费者的身份的信息以及描述那些消费者被授权相对于各种系统资源(例如，文件、目录、应用、通信端口，存储器片段，等等)执行的动作的信息。身份管理模块1028还可以包括关于每个消费者的描述信息以及关于描述信息可以如何以及被谁访问和修改的管理。

[0126] 图11示出了本发明的各种实施例可以在其中实现的示例计算机系统1100。系统1100可以被用来实现上述任何计算机系统。如图中所示，计算机系统1100包括经由总线子系统1102与多个外围子系统通信的处理单元1104。这些外围子系统可以包括处理加速单元1106、I/O子系统1108、存储子系统1118和通信子系统1124。存储子系统1118包括有形的计算机可读存储介质1122和系统存储器1110。

[0127] 总线子系统1102提供了用于让计算机系统1100的各种部件和子系统按预期彼此通信的机制。虽然总线子系统1102被示意性地示为单条总线，但是总线子系统的备选实施例可以利用多条总线。总线子系统1102可以是几种类型的总线结构中的任意一种，包括存储器总线或存储器控制器、外围总线以及使用任何各种总线体系架构的局部总线。例如，这种体系架构可以包括工业标准体系架构(ISA)总线、微通道体系架构(MCA)总线、增强型ISA(EISA)总线、视频电子标准协会(VESA)局部总线，以及外围部件互连(PCI)总线，该总线可以被实现为按IEEE P1386.1标准制造的Mezzanine总线。

[0128] 可以被实现为一个或多个集成电路(例如，常规的微处理器或微控制器)的处理单元1104控制计算机系统1100的操作。一个或多个处理器可以被包括在处理单元1104中。这些处理器可以包括单核或多核处理器。在某些实施例中，处理单元1104可以被实现为一个或多个独立的处理单元1132和/或1134，在每个处理单元中包括单核或多核处理器。在其它实施例中，处理单元1104也可以被实现为通过将两个双核处理器集成到单个芯片中所形成的四核处理单元。

[0129] 在各种实施例中，处理单元1104可以响应于程序代码而执行各种程序，并且可以

维护多个并发执行的程序或进程。在任何给定时间，要被执行的程序代码的一些或全部可以驻留在(一个或多个)处理器1104中和/或存储子系统1118中。通过适当的编程，(一个或多个)处理器1104可以提供上述各种功能。计算机系统1100还可以包括处理加速单元1106，该单元可以包括数字信号处理器(DSP)、专用处理器等等。

[0130] I/O子系统1108可以包括用户接口输入设备和用户接口输出设备。用户接口输入设备可以包括键盘、诸如鼠标或轨迹球的定点设备、结合到显示器中的触摸板或触摸屏、滚轮、点击轮、拨盘、按钮、开关、小键盘、具有语音命令识别系统的音频输入设备、麦克风以及其它类型的输入设备。用户接口输入设备可以包括，例如，运动感测和/或手势识别设备，诸如使得用户能够控制输入设备并与之交互的Microsoft**Kinect®**运动传感器，诸如通过利用手势和口头命令的自然用户接口的Microsoft**Xbox®**360游戏控制器。用户接口输入设备还可以包括眼睛姿势识别设备，诸如检测来自用户的眼睛活动(例如，在拍照和/或进行菜单选择的“眨眼”)并且将眼睛姿势作为输入变换到输入设备(例如，Google**Glass®**)的Google**Glass®**眨眼检测器。此外，用户接口输入设备可以包括使用户能够通过语音命令与语音识别系统(例如，**Siri®**·导航器)进行交互的语音识别感测设备。

[0131] 用户接口输入设备还可以包括但不限于，三维(3D)鼠标、操纵杆或指向棒、游戏手柄和绘图板，以及音频/视频设备，诸如扬声器、数码相机、数码摄像机、便携式媒体播放器、网络摄像头、图像扫描仪、指纹扫描仪、条形码阅读器的三维扫描仪、三维打印机、激光测距仪和视线跟踪设备。此外，用户接口输入设备可以包括，例如，医学成像输入设备，诸如计算机断层扫描、磁共振成像、正电子发射断层摄影术、医疗超声设备。用户接口输入设备还可以包括，例如，诸如MIDI键盘、数字乐器等的音频输入设备。

[0132] 用户接口输出设备可以包括显示子系统、指示灯，或者诸如音频输出设备的非视觉显示器等等。显示子系统可以是阴极射线管(CRT)、诸如使用液晶显示器(LCD)或等离子显示器的平板设备、投影设备、触摸屏等等。一般而言，术语“输出设备”的使用意在包括用于从计算机系统1100向用户或其它计算机输出信息的所有可能类型的设备和机制。例如，用户接口输出设备可以包括但不限于，在视觉上传达文本、图形和音频/视频信息的各种显示设备，诸如监视器、打印机、扬声器、耳机、汽车导航系统、绘图仪、语音输出设备，以及调制解调器。

[0133] 计算机系统1100可以包括包含软件元件的存储子系统1118，示为当前位于系统存储器1110内。系统存储器1110可以存储可以加载并可在处理单元1104上执行的程序指令，以及在这些程序执行期间所生成的数据。

[0134] 取决于计算机系统1100的配置和类型，系统存储器1110可以是易失性的(诸如随机存取存储器(RAM))和/或非易失性的(诸如只读存储器(ROM)、闪存存储器等等)。RAM通常包含可以让处理单元1104立即访问和/或目前正被处理单元1104操作和执行的数据和/或程序模块。在一些实现中，系统存储器1110可以包括多种不同类型的存储器，例如静态随机存取存储器(SRAM)或动态随机存取存储器(DRAM)。在一些实现中，包含有助于在计算机系统1100内的元件之间传送信息的基本例程的基本输入/输出系统(BIOS)，诸如在启动期间通常可以存储在ROM中。作为例子，但不是限制，系统存储器1110还说明了可以包括客户端应用、web浏览器、中间层应用、关系数据库管理系统(RDBMS)等的应用程序1112，程序数据1114，以及操作系统1116。作为例子，操作系统1116可以包括各种版本的Microsoft

Windows®,AppleMacintosh®和/或Linux操作系统,各种市售UNIX®或类UNIX操作系统(包括但不限于各种GNU/Linux操作系统、GoogleChrome®操作系统等等)和/或诸如iOS、Windows®Phone、Android®OS、BlackBerry®10OS和Palm®OS的移动操作系统。

[0135] 存储子系统1118还可以提供用于存储提供一些实施例的功能的基本编程和数据结构的有形的计算机可读存储介质。当由处理器执行时提供上述功能的软件(程序、代码模块、指令)可以被存储在存储子系统1118中。这些软件模块或指令可以由处理单元1104执行。存储子系统1118还可以提供用于存储根据本发明被使用的数据的储存库。

[0136] 存储子系统1100还可以包括可进一步被连接到计算机可读存储介质1122的计算机可读存储介质读取器1120。与系统存储器1110一起并且,可选地与其相结合,计算机可读存储介质1122可以全面表示用于临时和/或更持久地包含、存储、传送和检索计算机可读信息的远程、本地、固定和/或可移动存储设备加存储介质。

[0137] 包含代码或代码的各部分的计算机可读存储介质1122还可以包括本领域中已知或使用的任何适当介质,包括存储介质和通信介质,诸如但不限于,以用于信息存储和/或传输的任何方法或技术实现的易失性和非易失性、可移动和不可移动介质。这可以包括有形的计算机可读存储介质,诸如RAM、ROM、电可擦除可编程ROM(EEPROM)、闪存存储器或其它存储器技术、CD-ROM、数字多功能盘(DVD)或其它光学储存器、磁带盒、磁带、磁盘储存器或其它磁存储设备,或者其它有形的计算机可读介质。这还可以包括非有形的计算机可读介质,诸如数据信号、数据传输,或者可以被用来传送期望信息并且可以由计算系统1100访问的任何其它介质。

[0138] 作为例子,计算机可读存储介质1122可以包括从不可移动的、非易失性磁介质读取或写到其的硬盘驱动器、从可移动的、非易失性磁盘读取或写到其的磁盘驱动器,以及从可移动的、非易失性光盘(诸如CD ROM、DVD和Blu-Ray®盘或其它光学介质)读取或写到其的光盘驱动器。计算机可读存储介质1122可以包括但不限于,Zip®驱动器、闪存存储器卡、通用串行总线(USB)闪存驱动器、安全数字(SD)卡、DVD盘、数字录像带等等。计算机可读存储介质1122还可以包括基于非易失性存储器的固态驱动器(SSD),诸如基于闪存存储器的SSD、企业闪存驱动器、固态ROM等,基于易失性存储器的SSD,诸如固态RAM、动态RAM、静态RAM,基于DRAM的SSD,磁阻RAM(MRAM)SSD,以及使用基于DRAM和闪存存储器的SSD的组合的混合SSD。盘驱动器及其相关联的计算机可读介质可以为计算机系统1100提供计算机可读指令、数据结构、程序模块及其它数据的非易失性存储。

[0139] 通信子系统1124提供到其它计算机系统和网络的接口。通信子系统1124充当用于从其它系统接收数据和从计算机系统1100向其它系统发送数据的接口。例如,通信子系统1124可以使计算机系统1100能够经由互联网连接到一个或多个设备。在一些实施例中,通信子系统1124可以包括用于访问无线语音和/或数据网络的射频(RF)收发器部件(例如,利用蜂窝电话技术,诸如3G、4G或EDGE(增强型数据速率全球演进)的先进数据网络技术,Wi-Fi(IEEE 802.11系列标准,或其它移动通信技术,或它们的任意组合))、全球定位系统(GPS)接收器部件,和/或其它部件。在一些实施例中,除了无线接口或代替其,通信子系统1124还可以提供有线网络连接(例如,以太网)。

[0140] 在一些实施例中，通信子系统1124还可以代表可以使用计算机系统1100的一个或多个用户接收结构化和/或非结构化数据馈送1126、事件流1128、事件更新1130等形式的输入通信。

[0141] 作为例子，通信子系统1124可以被配置为实时地从社交网络和/或其它通信服务的用户接收数据馈送1126，诸如Twitter®馈送、Facebook®更新、诸如丰富站点摘要(RSS)馈送的web馈送和/或来自一个或多个第三方信息源的实时更新。

[0142] 此外，通信子系统1124还可以被配置为接收连续数据流形式的数据，这可以包括实时事件的事件流1128和/或事件更新1130，其本质可以是连续的或无界的，没有明确的结束。生成连续数据的应用的例子可以包括，例如，传感器数据应用、金融股票、网络性能测量工具(例如，网络监视和流量管理应用)、点击流分析工具、汽车流量监视等等。通信子系统1124还可以被配置为向一个或多个数据库输出结构化和/或非结构化数据馈送1126、事件流1128、事件更新1130等等，这一个或多个数据库可以与耦合到计算机系统1100的一个或多个流数据源计算机进行通信。

[0143] 计算机系统1100可以是各种类型之一，包括手持便携式设备(例如，iPhone®蜂窝电话、iPad®计算平板、PDA)、可穿戴设备(例如，Google Glass®头戴式显示器)、PC、工作站、大型机、信息站，服务器机架，或任何其它数据处理系统。

[0144] 由于计算机和网络的不断变化的本质，图中所绘出的计算机系统1100的描述仅仅是作为具体的例子。可能具有比图中所绘的系统更多或更少部件的许多其它配置。例如，定制的硬件也可以被使用和/或特定的元件可以在硬件、固件、软件(包括小应用程序)或它们的组合中实现。另外，也可以采用到诸如网络输入/输出设备之类的其它计算设备的连接。基于本文提供的公开内容和示教，本领域普通技术人员将认识到实现各种实施例的其它方式和/或方法。

[0145] 在前述的说明书中，本发明的各方面参照其具体实施例进行了描述，但本领域技术人员将认识到，本发明不限于此。上述发明的各个特征和各方面可以单独或联合使用。另外，在不背离本说明书的更广泛精神和范围的情况下，实施例可以在除本文所述的那些之外的任何数目的环境和应用中被使用。因此，本说明书和附图应当被认为是说明性而不是限制性的。

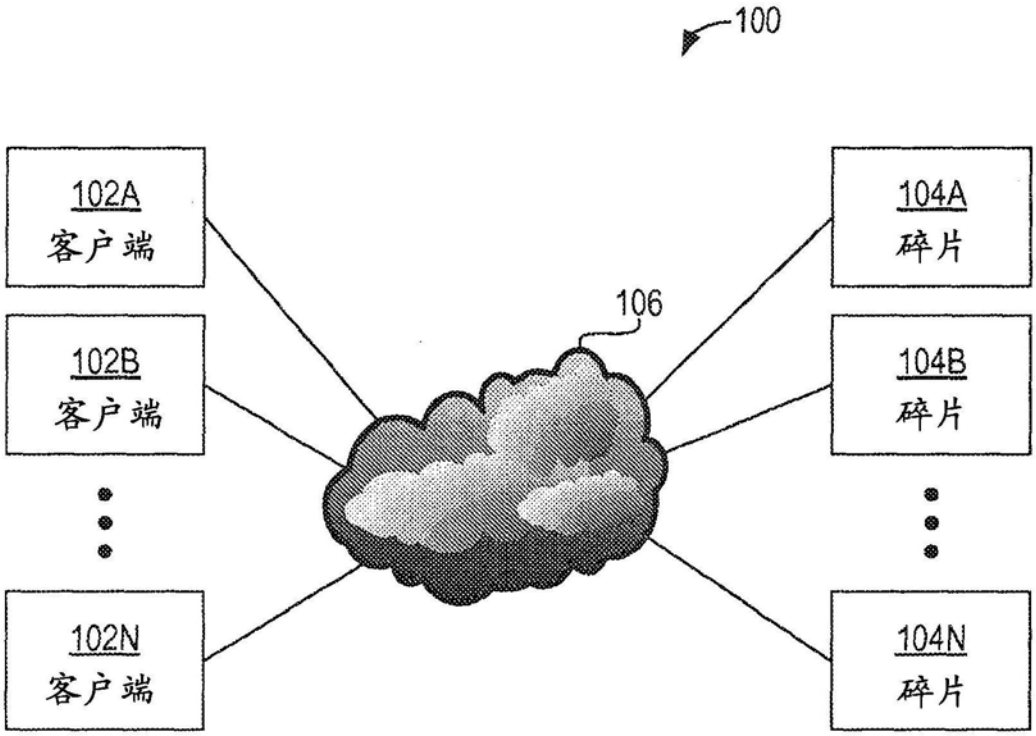

100
102A
客户端
102B
客户端
102N
客户端
106
104A
碎片
104B
碎片
104N
碎片

图1

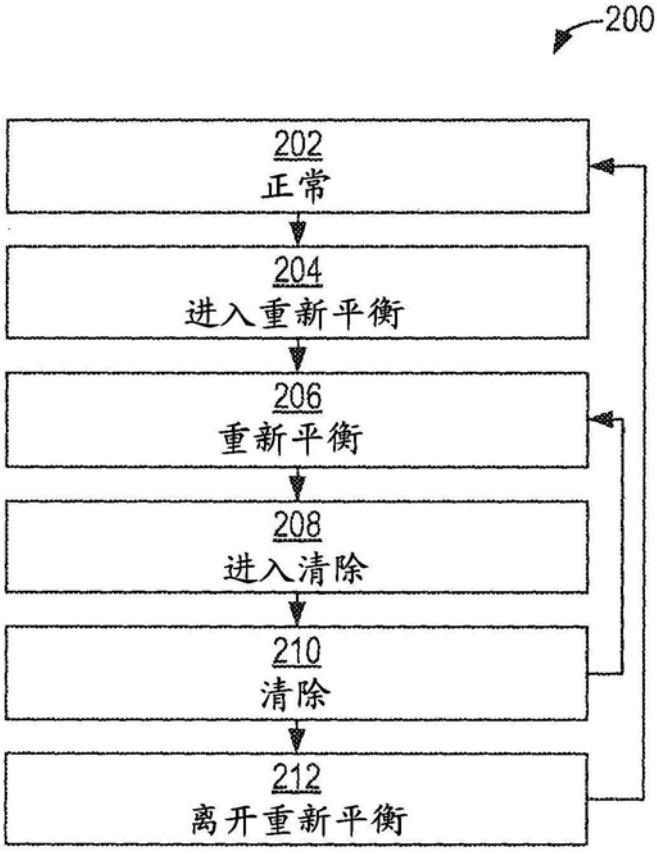

200
202
正常
204
进入重新平衡
206
重新平衡
208
进入清除
210
清除
212
离开重新平衡

图2

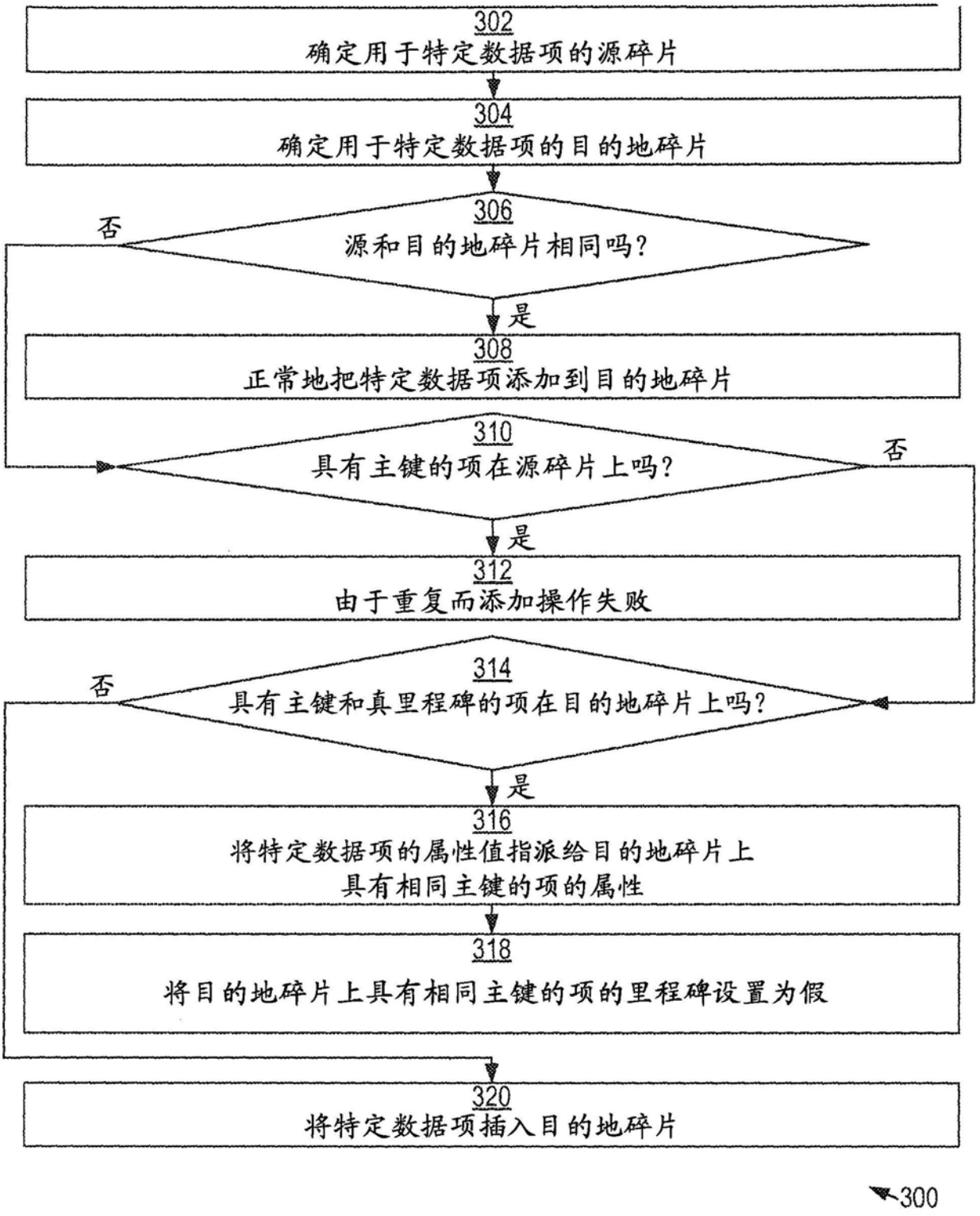

302
确定用于特定数据项的源碎片
304
确定用于特定数据项的目的地碎片
306
源和目的地碎片相同吗?
否
是
308
正常地把特定数据项添加到目的地碎片
310
具有主键的项在源碎片上吗?
否
是
312
由于重复而添加操作失败
314
具有主键和真里程碑的项在目的地碎片上吗?
否
是
316
将特定数据项的属性值指派给目的地碎片上
具有相同主键的项的属性
318
将目的地碎片上具有相同主键的项的里程碑设置为假
320
将特定数据项插入目的地碎片
300

图3

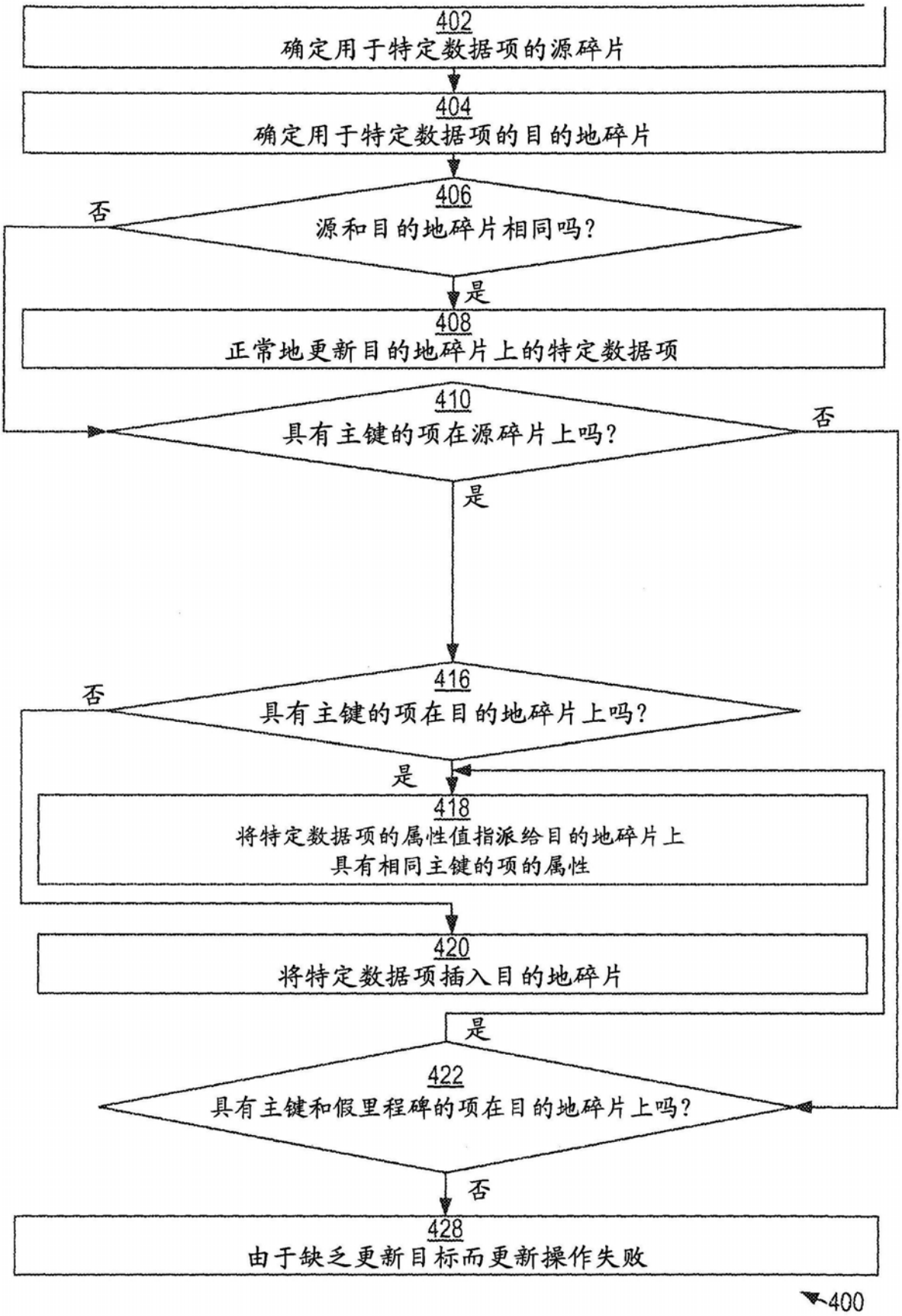

402
确定用于特定数据项的源碎片
404
确定用于特定数据项的目的地碎片
406
源和目的地碎片相同吗?
否
是
408
正常地更新目的地碎片上的特定数据项
410
具有主键的项在源碎片上吗?
否
是
416
具有主键的项在目的地碎片上吗?
否
是
418
将特定数据项的属性值指派给目的地碎片上
具有相同主键的项的属性
420
将特定数据项插入目的地碎片
是
422
具有主键和假里程碑的项在目的地碎片上吗?
否
428
由于缺乏更新目标而更新操作失败
400

图4

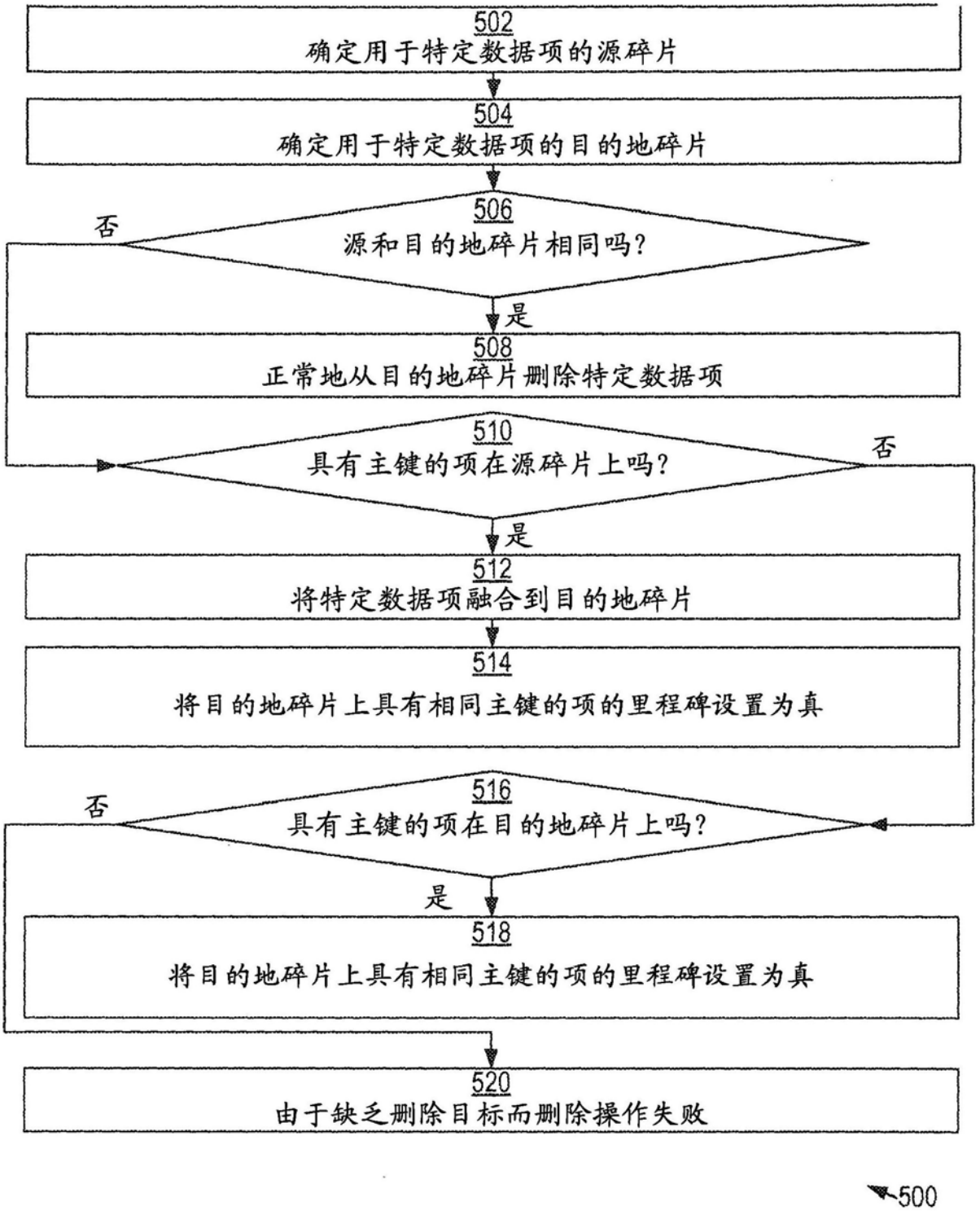

502
确定用于特定数据项的源碎片
504
确定用于特定数据项的目的地碎片
506
源和目的地碎片相同吗?
否
是
508
正常地从目的地碎片删除特定数据项
510
具有主键的项在源碎片上吗?
否
是
512
将特定数据项融合到目的地碎片
514
将目的地碎片上具有相同主键的项的里程碑设置为真
516
具有主键的项在目的地碎片上吗?
否
是
518
将目的地碎片上具有相同主键的项的里程碑设置为真
520
由于缺乏删除目标而删除操作失败
500

图5

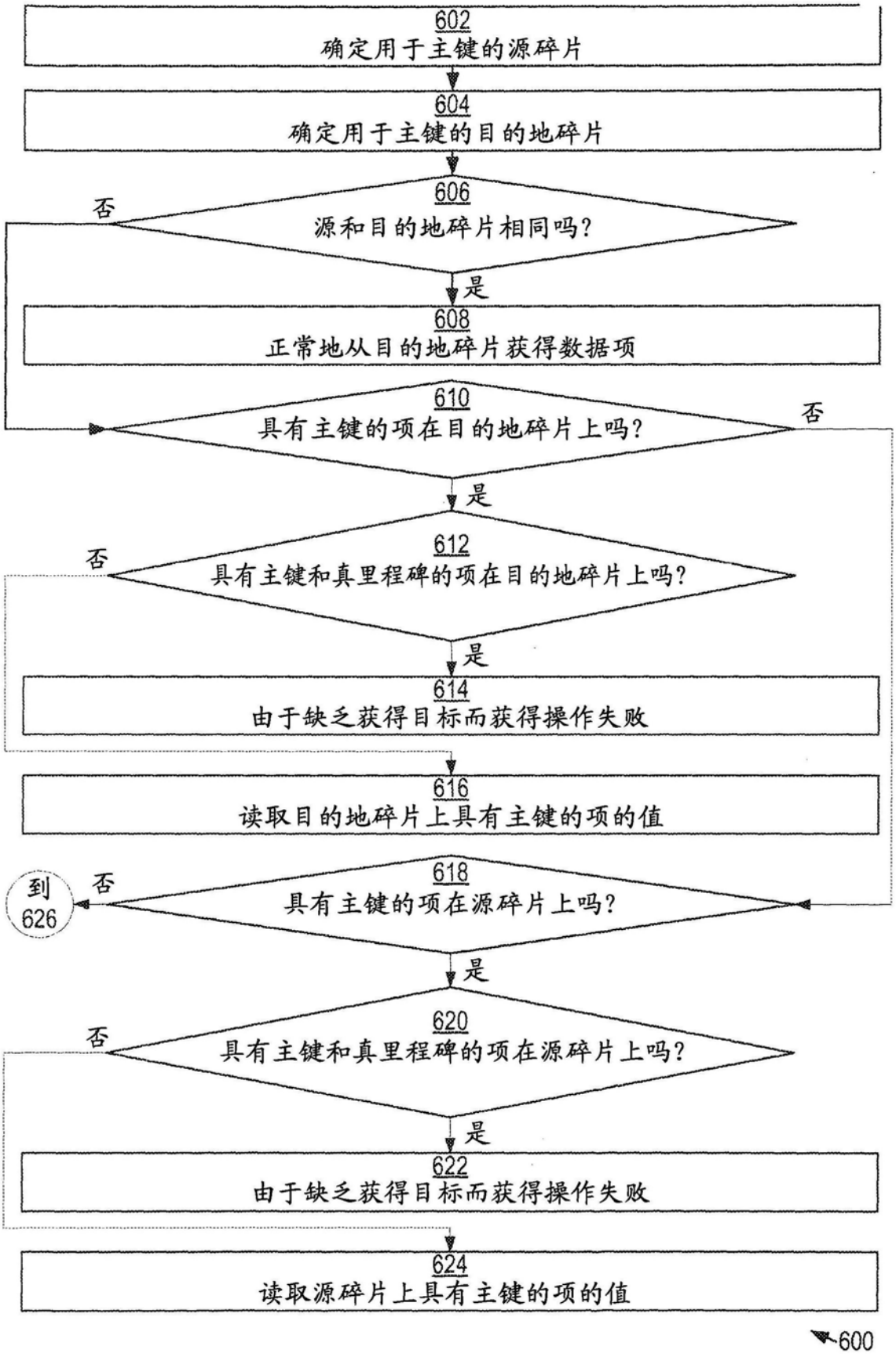

602
确定用于主键的源碎片
604
确定用于主键的目的地碎片
606
源和目的地碎片相同吗?
否
是
608
正常地从目的地碎片获得数据项
610
具有主键的项在目的地碎片上吗?
否
是
612
具有主键和真里程碑的项在目的地碎片上吗?
否
是
614
由于缺乏获得目标而获得操作失败
616
读取目的地碎片上具有主键的项的值
618
具有主键的项在源碎片上吗?
到626
否
是
620
具有主键和真里程碑的项在源碎片上吗?
否
是
622
由于缺乏获得目标而获得操作失败
624
读取源碎片上具有主键的项的值
600

图6A

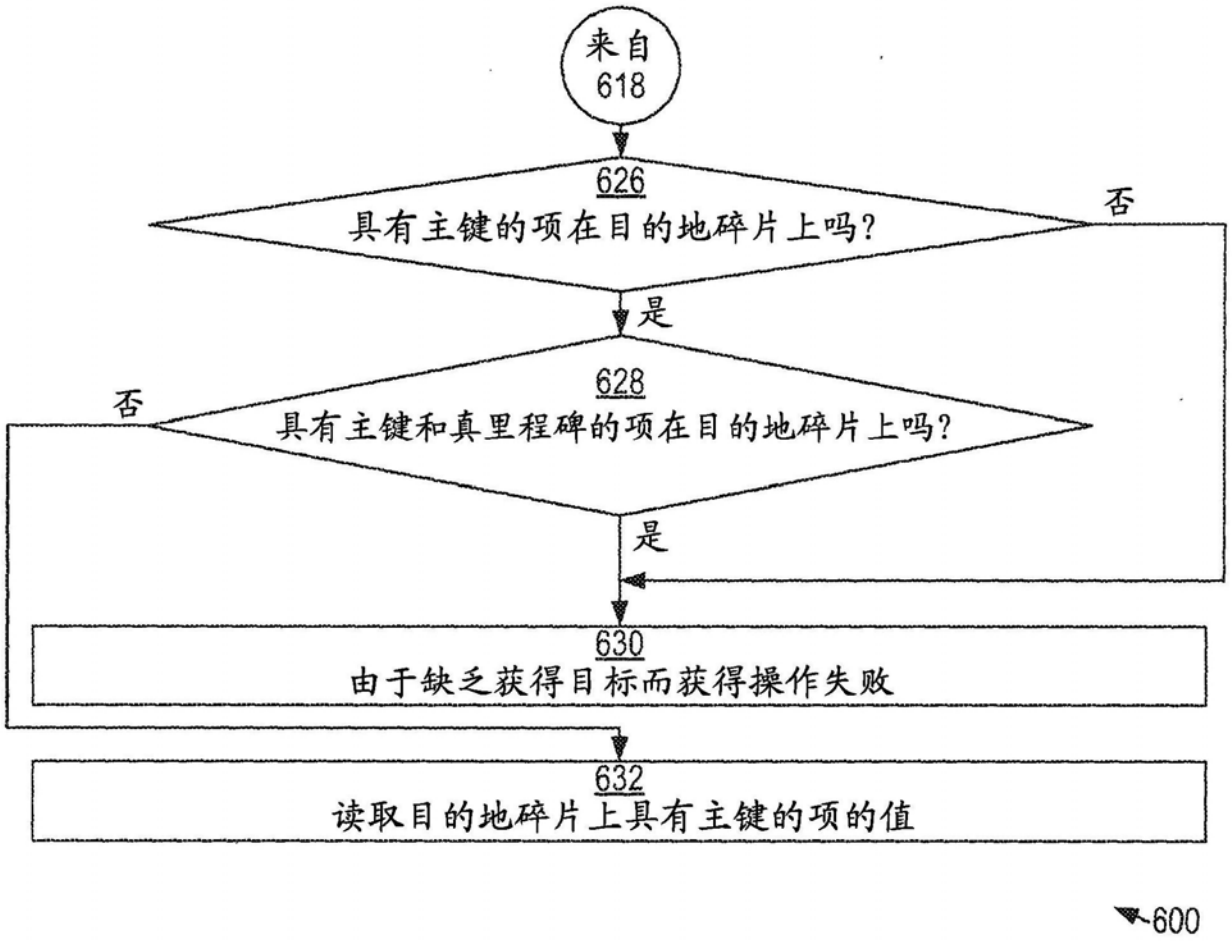

来自
618
626
具有主键的项在目的地碎片上吗?
否
是
628
具有主键和真里程碑的项在目的地碎片上吗?
否
是
630
由于缺乏获得目标而获得操作失败
632
读取目的地碎片上具有主键的项的值
600

图6B

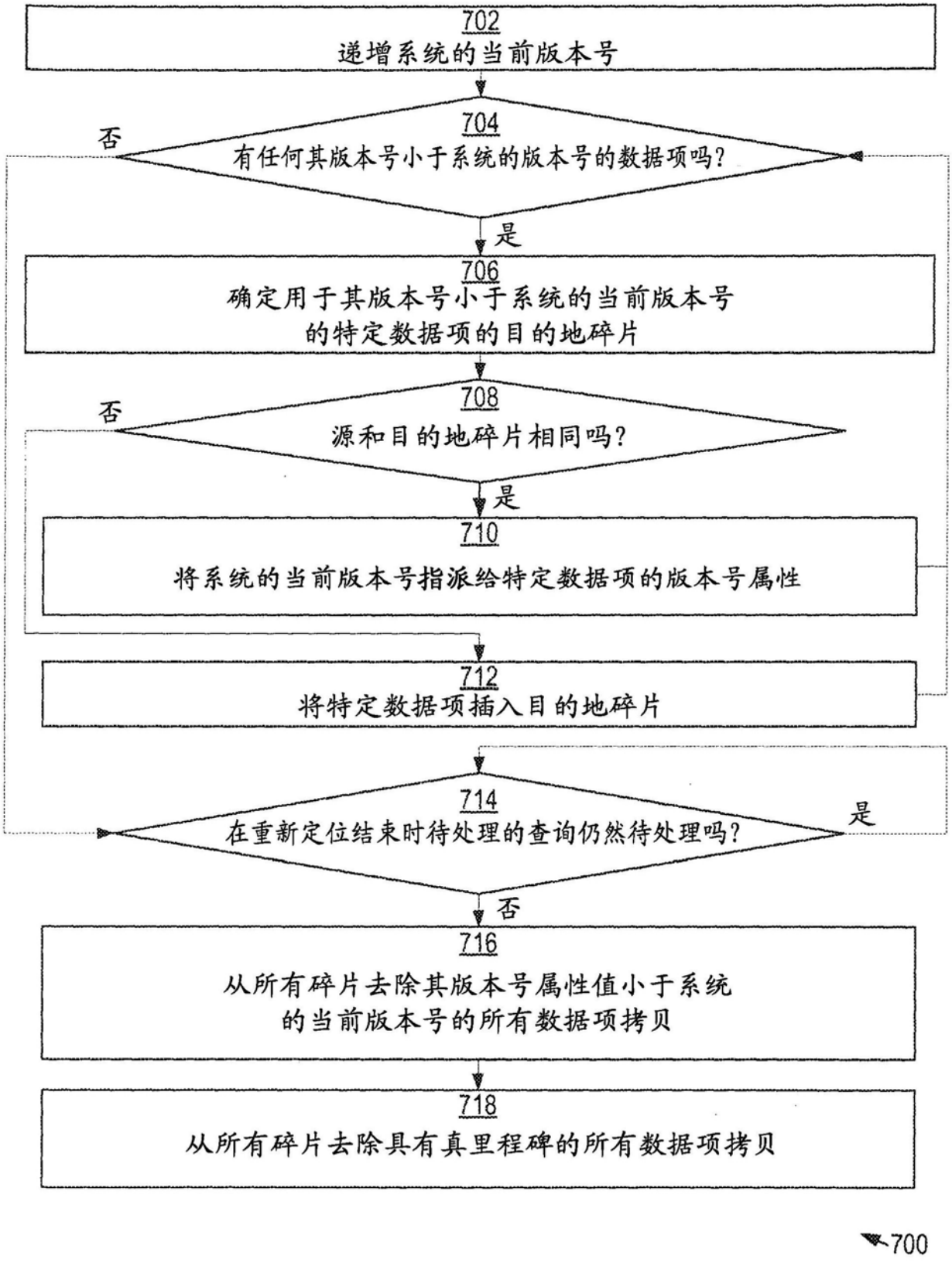

702
递增系统的当前版本号
704
有任何其版本号小于系统的版本号的数据项吗?
否
是
706
确定用于其版本号小于系统的当前版本号
的特定数据项的目的地碎片
708
源和目的地碎片相同吗?
否
是
710
将系统的当前版本号指派给特定数据项的版本号属性
712
将特定数据项插入目的地碎片
714
在重新定位结束时待处理的查询仍然待处理吗?
是
否
716
从所有碎片去除其版本号属性值小于系统
的当前版本号的所有数据项拷贝
718
从所有碎片去除具有真里程碑的所有数据项拷贝
700

图7

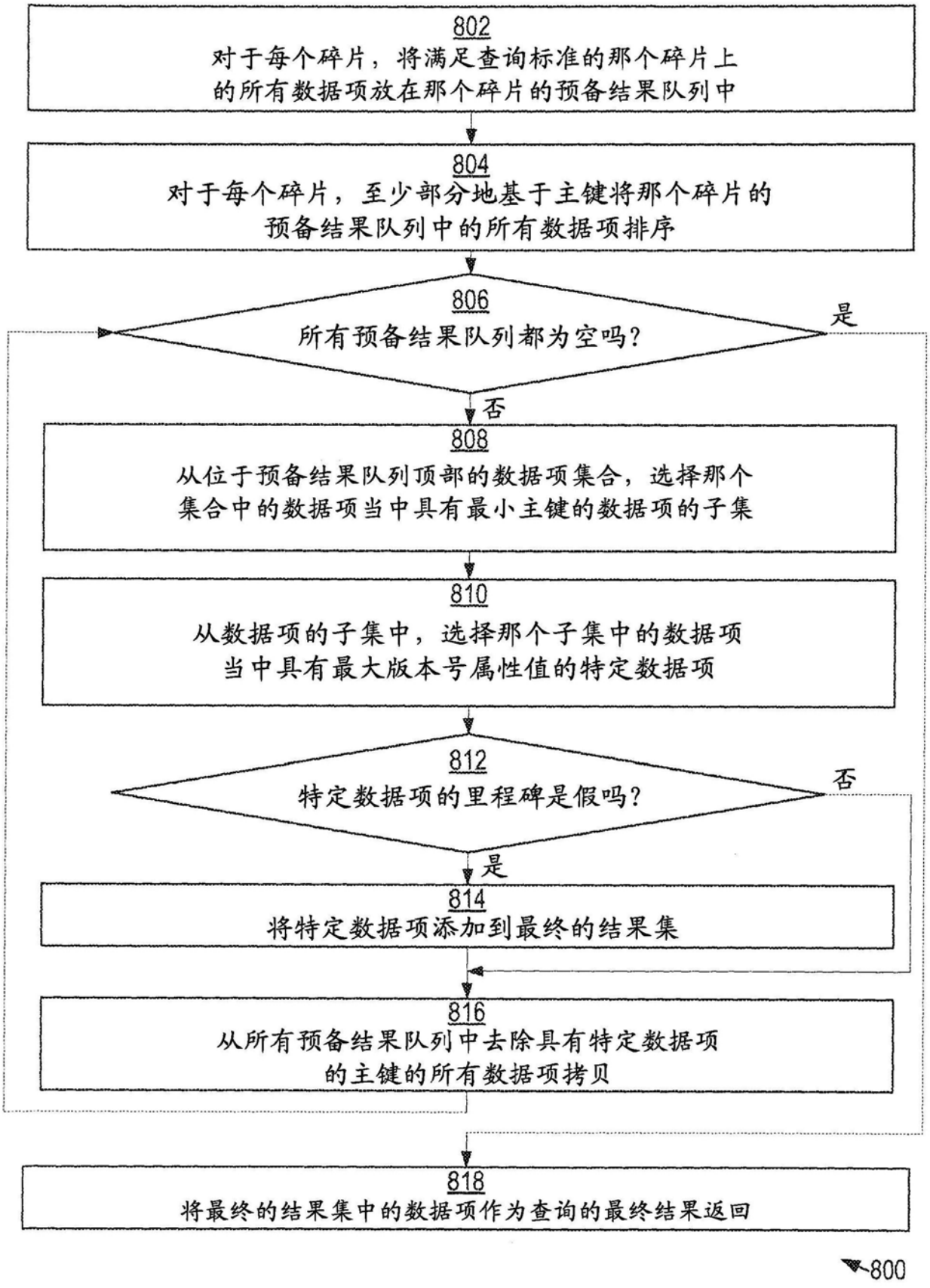

802
对于每个碎片，将满足查询标准的那个碎片上的所有数据项放在那个碎片的预备结果队列中
804
对于每个碎片，至少部分地基于主键将那个碎片的预备结果队列中的所有数据项排序
806
所有预备结果队列都为空吗？
是
否
808
从位于预备结果队列顶部的数据项集合，选择那个集合中的数据项当中具有最小主键的数据项的子集
810
从数据项的子集中，选择那个子集中的数据项当中具有最大版本号属性值的特定数据项
812
特定数据项的里程碑是假吗？
否
是
814
将特定数据项添加到最终的结果集
816
从所有预备结果队列中去除具有特定数据项的主键的所有数据项拷贝
818
将最终的结果集中的数据项作为查询的最终结果返回
800

图8

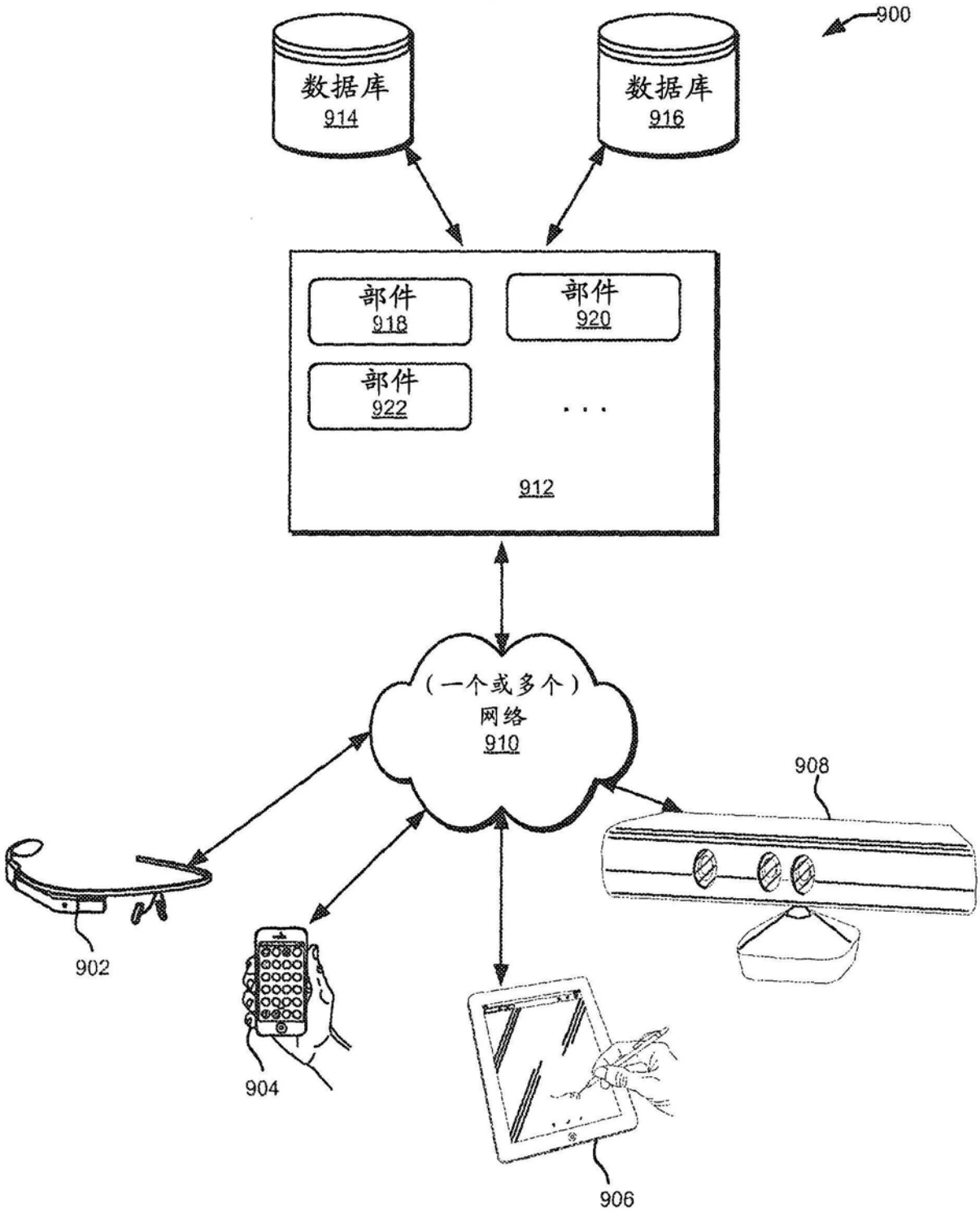

900
数据库
914
数据库
916
部件
918
部件
920
部件
922
912
（一个或多个）
网络
910
908
902
904
906

图9

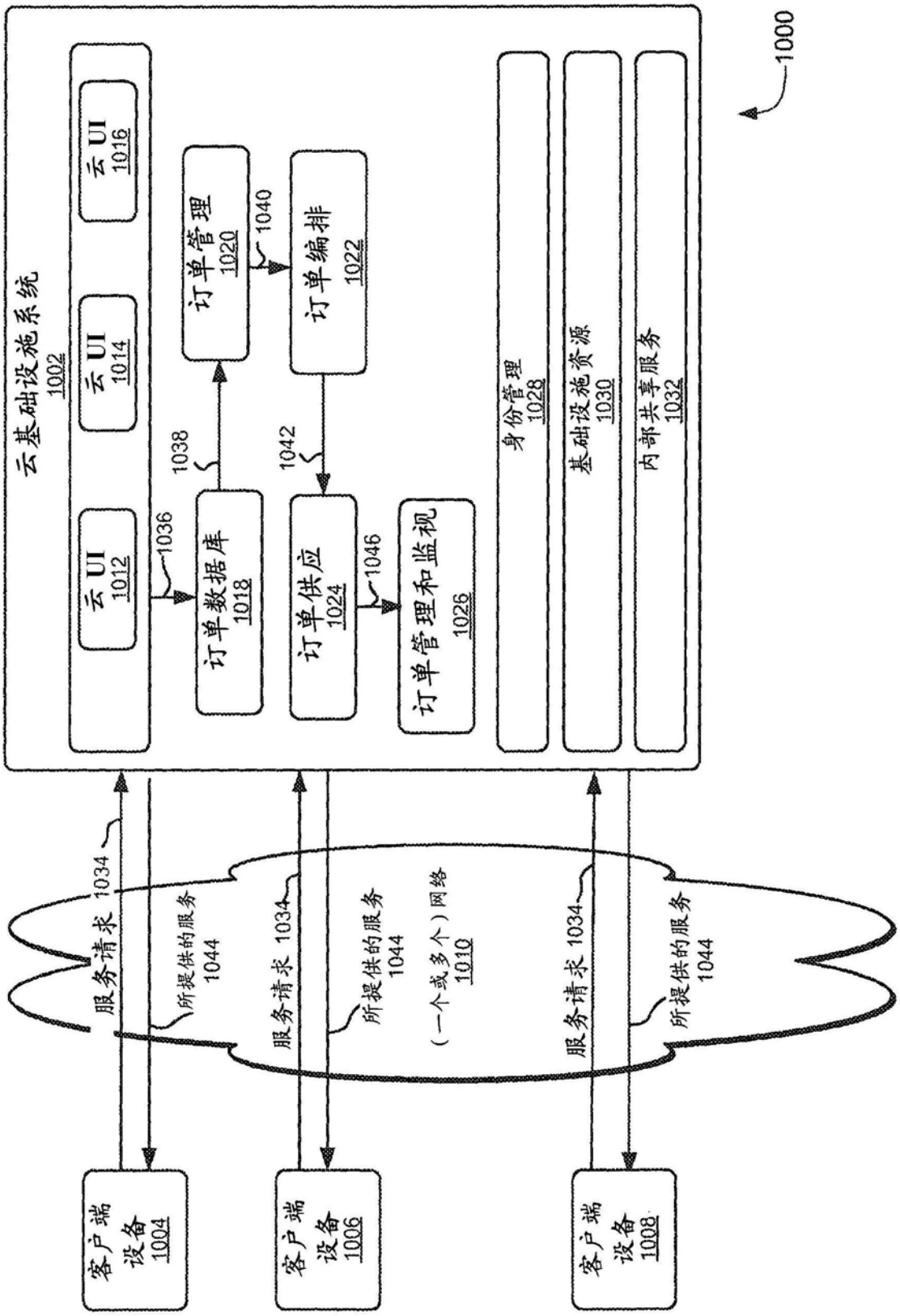

1000
云基础设施系统
1002
云 UI
1012
云 UI
1014
云 UI
1016
订单数据库
1018
订单管理
1020
订单供应
1024
订单编排
1022
订单管理和监视
1026
身份管理
1028
基础设施资源
1030
内部共享服务
1032
1036
1038
1040
1042
1046
服务请求 1034
所提供的服务
1044
（一个或多个）网络
1010
客户端
设备
1004
客户端
设备
1006
客户端
设备
1008

图10

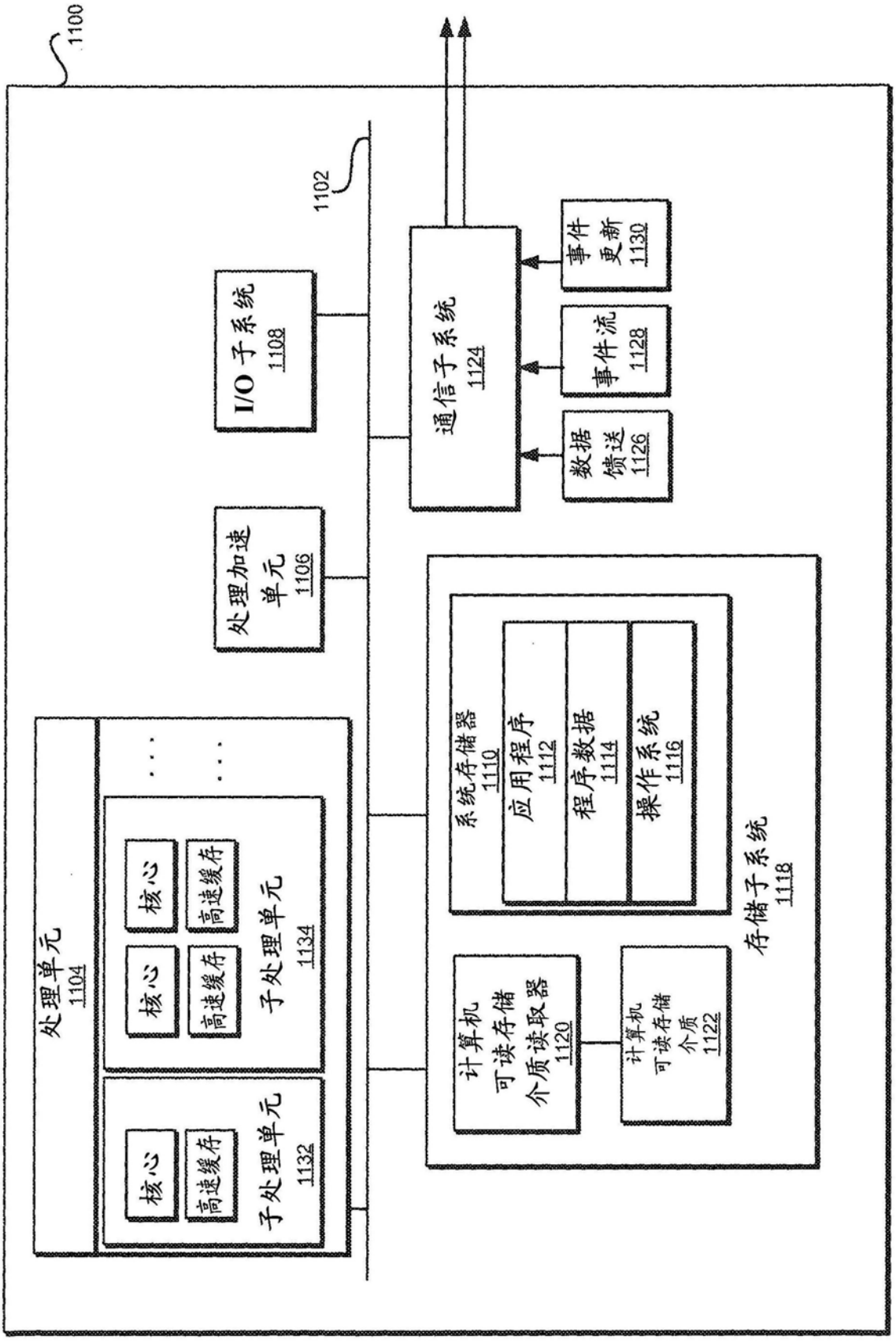

1100
1102
处理单元
1104
核心
高速缓存
子处理单元
1132
核心
高速缓存
核心
高速缓存
子处理单元
1134
I/O 子系统
1108
处理加速单元
1106
通信子系统
1124
数据馈送
1126
事件流
1128
事件更新
1130
存储子系统
1118
系统存储器
1110
应用程序
1112
程序数据
1114
操作系统
1116
计算机可读存储介质读取器
1120
计算机可读存储介质
1122

图11