(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 111247518 B
(45)授权公告日 2024.05.14

(21)申请号 201880068665.2
(22)申请日 2018.10.18
(65)同一申请的已公布的文献号
申请公布号 CN 111247518 A
(43)申请公布日 2020.06.05
(30)优先权数据
15/793,100 2017.10.25 US
15/813,577 2017.11.15 US
(85)PCT国际申请进入国家阶段日
2020.04.21
(86)PCT国际申请的申请数据
PCT/EP2018/078495 2018.10.18
(87)PCT国际申请的公布数据
WO2019/081322 EN 2019.05.02

(73)专利权人 国际商业机器公司
地址 美国纽约阿芒克
(72)发明人 C·N·小瓦伦 M·里安
(74)专利代理机构 北京市金杜律师事务所 11256
专利代理师 酆迅 姚杰

(51)Int.Cl.
G06F 16/22(2019.01)
G06F 16/24(2019.01)
(56)对比文件
CN 102473084 A,2012.05.23
CN 104115146 A,2014.10.22
US 2010312749 A1,2010.12.09
US 2016110391 A1,2016.04.21
审查员 刘剑

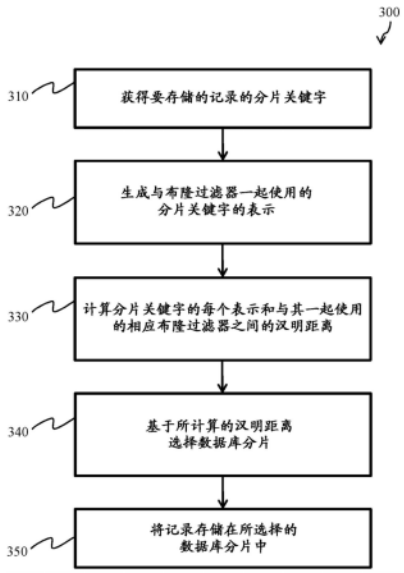
权利要求书5页 说明书14页 附图4页

(54)发明名称

用于数据库分片的方法和系统

(57)摘要

提供了用于在分片数据库的分片表中存储记录和从分片数据库的分片表检索记录的计算机实现的方法、计算机程序产品和系统。分片数据库包括多个数据库分片。数据库分片各自被配置为存储分片表的记录的子集。数据库分片还各自与相应的布隆过滤器相关联。通过生成与多个相应布隆过滤器中的每一个一起使用的要存储的记录的分片关键字的相应表示,将该记录插入到分片表中。然后计算多个布隆过滤器中的每一个布隆过滤器的当前值和与该布隆过滤器一起使用的分片关键字的相应表示之间的汉明距离,并且基于所计算的汉明距离从那些数据库分片中选择用于存储该记录的数据库分片。



CN 111247518 B

1.一种用于将记录存储在分片数据库的分片表中的计算机实现的方法,所述分片数据库包括多个数据库分片,所述多个数据库分片中的每个数据库分片被配置为存储所述分片表的记录的子集并且与相应的布隆过滤器相关联,所述布隆过滤器被配置为提供关于与特定分片关键字相关联的记录是否已经被存储在所述数据库分片中的指示,所述计算机实现的方法包括:

获得要被存储的所述记录的分片关键字,所述分片关键字基于所述记录的一个或多个字段;

针对多个相应布隆过滤器中的每一个布隆过滤器,通过将所述布隆过滤器所使用的一个或多个散列函数应用于所述分片关键字来生成与所述布隆过滤器一起使用的所述分片关键字的相应表示;

计算所述多个布隆过滤器中的每一个布隆过滤器的当前值和与所述布隆过滤器一起使用的所述分片关键字的所述相应表示之间的汉明距离;

基于所计算的汉明距离来选择数据库分片;以及

将所述记录存储在所选择的数据库分片中。

2.根据权利要求1所述的计算机实现的方法,其中所述相应布隆过滤器中的每一个布隆过滤器使用相同数量的比特和相同的散列函数。

3.根据权利要求1所述的计算机实现的方法,其中所述相应布隆过滤器中的每一个布隆过滤器是计数过滤器。

4.根据权利要求1所述的计算机实现的方法,其中选择数据库分片包括选择与所计算的汉明距离是所计算的汉明距离之中的最小值的相应布隆过滤器相关联的数据库分片。

5.根据权利要求4所述的计算机实现的方法,其中选择与所计算的汉明距离是所计算的汉明距离之中的最小值的相应布隆过滤器相关联的数据库分片,包括从与所计算的汉明距离是最小值的相应布隆过滤器相关联的那些数据库分片之中随机选择所述数据库分片。

6.根据权利要求1所述的计算机实现的方法,其中所述分片数据库还包括每个数据库分片可用于存储新记录或者不可用于存储新记录的指示,并且其中为其生成所述分片关键字的相应表示的所述多个布隆过滤器不包括与被指示为不可用的数据库分片相关联的任何布隆过滤器。

7.根据权利要求6所述的计算机实现的方法,其中为其生成所述分片关键字的表示的所述多个布隆过滤器包括与被指示为可用的所有所述数据库分片相关联的相应布隆过滤器。

8.根据权利要求1所述的计算机实现的方法,其中所述分片数据库还包括用于对已经插入到所述数据库分片中的每个数据库分片中的记录的相应数量进行计数的相应插入计数器,并且其中所述计算机实现的方法还包括:

针对所选择的数据库分片递增所述插入计数器。

9.根据权利要求8所述的计算机实现的方法,其中所述计算机实现的方法还包括:

确定已被插入到所选择的数据库分片中的记录的数量是否超过预定阈值;以及

响应于确定已被插入到所选择的数据库分片中的所述记录的数量超过所述预定阈值,将所选择的数据库分片标记为不可用于存储新记录。

10.根据权利要求1所述的计算机实现的方法,其中所述计算机实现的方法还包括:

将新数据库分片和相关联的布隆过滤器添加到所述分片数据库。

11.一种用于从分片数据库的分片表检索记录的计算机实现的方法,所述分片数据库包括多个数据库分片,所述多个数据库分片中的每个数据库分片被配置为存储所述分片表的记录的子集并且与相应的布隆过滤器相关联,所述布隆过滤器被配置为提供关于与特定分片关键字相关联的记录是否已经被存储在所述数据库分片中的指示,所述计算机实现的方法包括:

获得用于要检索的所述记录的分片关键字,所述分片关键字基于所述记录的一个或多个字段;

针对与所述多个数据库分片中的每个数据库分片相关联的相应布隆过滤器中的每个布隆过滤器,通过将由所述布隆过滤器使用的一个或多个散列函数应用于所述分片关键字来生成与所述布隆过滤器一起使用的所述分片关键字的相应表示;

标识所述相应布隆过滤器中的哪些布隆过滤器与其的所述分片关键字的相应表示匹配;

搜索与所标识的布隆过滤器相关联的所述数据库分片以找到所述记录,其中所述记录被存储在基于计算的针对每个布隆过滤器的值和与所述布隆过滤器一起使用的所述分片关键字的所述相应表示之间的汉明距离而选择的数据库分片中;以及

提供所述记录。

12.根据权利要求11所述的计算机实现的方法,其中所述相应布隆过滤器中的每一个使用相同数量的比特和相同的散列函数。

13.一种用于存储分片数据库的系统,所述分片数据库包括多个数据库分片,所述多个数据库分片中的每个数据库分片被配置为存储分片表的记录的子集并且与相应的布隆过滤器相关联,所述布隆过滤器被配置为提供关于与特定分片关键字相关联的记录是否已经被存储在所述数据库分片中的指示,所述系统包括:

多个计算系统,每个计算系统被配置为存储所述多个数据库分片中的一个或多个;以及

控制器,其中所述控制器被配置为通过以下操作从所述分片数据库的分片表检索记录:

获得用于要检索的所述记录的分片关键字,所述分片关键字基于所述记录的一个或多个字段;

针对与所述多个数据库分片中的每个数据库分片相关联的相应布隆过滤器中的每个布隆过滤器,通过将由所述布隆过滤器使用的一个或多个散列函数应用于所述分片关键字来生成与所述布隆过滤器一起使用的所述分片关键字的相应表示;

标识所述相应布隆过滤器中的哪些布隆过滤器与其的所述分片关键字的相应表示匹配;

搜索与所标识的布隆过滤器相关联的数据库分片以找到所述记录,其中所述记录被存储在基于计算的针对每个布隆过滤器的值和与所述布隆过滤器一起使用的所述分片关键字的所述相应表示之间的汉明距离而选择的数据库分片中;以及

提供所述记录。

14.根据权利要求13所述的系统,其中,所述相应布隆过滤器中的每一个布隆过滤器使

用相同数量的比特和相同的散列函数。

15.根据权利要求13所述的系统,其中,所述相应布隆过滤器中的每一个布隆过滤器是计数过滤器。

16.根据权利要求13所述的系统,其中所述控制器还被配置为通过以下操作将记录存储在所述分片数据库的所述分片表中:

获得要被存储的所述记录的分片关键字,所述分片关键字基于所述记录的一个或多个字段;

针对多个相应布隆过滤器中的每一个布隆过滤器,通过将所述布隆过滤器所使用的一个或多个散列函数应用于所述分片关键字来生成与所述布隆过滤器一起使用的分片关键字的相应表示;

计算所述多个布隆过滤器中的每一个布隆过滤器的当前值和与所述布隆过滤器一起使用的分片关键字的相应表示之间的汉明距离;

基于所计算的汉明距离来选择数据库分片;以及

将所述记录存储在所选择的数据库分片中。

17.根据权利要求16所述的系统,其中选择数据库分片包括选择与所计算汉明距离是所计算汉明距离之中的最小值的相应布隆过滤器相关联的数据库分片。

18.根据权利要求16所述的系统,其中所述分片数据库还包括每个数据库分片是否可用于存储新记录的指示,并且其中当存储所述记录时为其生成所述分片关键字的相应表示的所述多个布隆过滤器不包括与被指示为不可用的数据库分片相关联的任何布隆过滤器。

19.根据权利要求16所述的系统,其中所述分片数据库还包括用于对已经插入到所述数据库分片中的每个数据库分片中的记录的相应数量进行计数的相应插入计数器,并且其中所述控制器还被配置为当所述记录被存储时递增针对所选择的数据库分片的所述插入计数器。

20.根据权利要求19所述的系统,其中,所述控制器还被配置为:

确定已被插入到所选择的数据库分片中的记录的所述数量是否超过预定阈值;以及

响应于确定已被插入到所选择的数据库分片中的记录的所述数量超过所述预定阈值,将所选择的数据库分片标记为不可用于存储新记录。

21.根据权利要求16所述的系统,其中,所述控制器还被配置为:

将新数据库分片和相关联的布隆过滤器添加到所述分片数据库。

22.根据权利要求16所述的系统,其中所述分片数据库还包括相应删除计数器,所述相应删除计数器用于对已经从所述多个数据库分片中的每个数据库分片删除的记录的相应数量进行计数。

23.根据权利要求22所述的系统,其中,所述控制器还被配置为:

确定已经从特定数据库分片删除的记录的数量是否超过所述数据库分片的相应预定阈值;以及响应于确定已经从所述特定数据库分片删除的记录的数目超过所述数据库分片的所述相应预定阈值,基于当前存储在所述数据库分片中的记录来重建所述数据库分片的布隆过滤器。

24.一种用于将记录存储在分片数据库的分片表中的计算机可读存储介质,所述分片数据库包括多个数据库分片,所述多个数据库分片中的每个数据库分片被配置为存储所述

分片表的记录的子集并且与相应的布隆过滤器相关联,所述布隆过滤器被配置为提供关于与特定分片关键字相关联的记录是否已经被存储在所述数据库分片中的指示,所述计算机可读存储介质具有编码在其上的:

第一程序指令,所述第一程序指令可由处理器执行以使所述处理器获得要被存储的所述记录的分片关键字,所述分片关键字基于所述记录的一个或多个字段;

第二程序指令,所述第二程序指令可由所述处理器执行以使所述处理器针对多个相应的所述布隆过滤器中的每一个布隆过滤器,通过将由所述布隆过滤器使用的一个或多个散列函数应用于所述分片关键字来生成与所述布隆过滤器一起使用的所述分片关键字的相应表示;

第三程序指令,所述第三程序指令可由所述处理器执行以使所述处理器计算所述多个布隆过滤器中的每一个布隆过滤器的当前值和与该布隆过滤器一起使用的所述分片关键字的相应表示之间的汉明距离;

第四程序指令,所述第四程序指令可由所述处理器执行以使所述处理器基于所计算的汉明距离来选择数据库分片;以及

第五程序指令,所述第五程序指令可由所述处理器执行以使所述处理器将所述记录存储在所选择的数据库分片中。

25.一种用于从分片数据库的分片表检索记录的计算机可读存储介质,所述分片数据库包括多个数据库分片,所述多个数据库分片中的每个数据库分片被配置为存储所述分片表的记录的子集并且与相应的布隆过滤器相关联,所述布隆过滤器被配置为提供关于与特定分片关键字相关联的记录是否已经被存储在所述数据库分片中的指示,所述计算机可读存储介质具有编码在其上的:

第一程序指令,所述第一程序指令可由处理器执行以使所述处理器获得用于要被检索的所述记录的分片关键字,所述分片关键字基于所述记录的一个或多个字段;

第二程序指令,所述第二程序指令可由所述处理器执行以使所述处理器针对与所述多个数据库分片中的每个数据库分片相关联的相应布隆过滤器中的每个布隆过滤器,通过将由所述布隆过滤器使用的一个或多个散列函数应用于所述分片关键字来生成和所述布隆过滤器一起使用的所述分片关键字的相应表示;

第三程序指令,所述第三程序指令可由所述处理器执行以使所述处理器识别所述相应布隆过滤器中的哪些布隆过滤器与其的所述分片关键字的相应的表示匹配;

第四程序指令,所述第四程序指令可由所述处理器执行以使所述处理器搜索与所识别的布隆过滤器相关联的所述数据库分片以找到所述记录,其中所述记录被存储在基于计算的针对每个布隆过滤器的值和与所述布隆过滤器一起使用的所述分片关键字的所述相应表示之间的汉明距离而选择的数据库分片中;以及

第五程序指令,所述第五程序指令可由所述处理器执行以使所述处理器提供所述记录。

26.一种用于将记录存储在分片数据库的分片表中的系统,所述分片数据库包括多个数据库分片,所述多个数据库分片中的每个数据库分片被配置为存储所述分片表的记录的子集并且与相应的布隆过滤器相关联,所述布隆过滤器被配置为提供关于与特定分片关键字相关联的记录是否已经被存储在所述数据库分片中的指示,所述系统可配置成:

获得用于要被存储的所述记录的分片关键字,所述分片关键字基于所述记录的一个或多个字段;

针对多个相应布隆过滤器中的每一个布隆过滤器,通过将所述布隆过滤器所使用的一个或多个散列函数应用于所述分片关键字来生成与所述布隆过滤器一起使用的所述分片关键字的相应表示;

计算所述多个布隆过滤器中的每一个布隆过滤器的当前值和与所述布隆过滤器一起使用的所述分片关键字的相应表示之间的汉明距离;

基于所计算的汉明距离来选择数据库分片;以及

将所述记录存储在所选择的数据库分片中。

27.一种用于从分片数据库的分片表检索记录的系统,所述分片数据库包括多个数据库分片,所述多个数据库分片中的每个数据库分片被配置为存储所述分片表的记录的子集并且与相应的布隆过滤器相关联,所述布隆过滤器被配置为提供关于与特定分片关键字相关联的记录是否已经被存储在所述数据库分片中的指示,所述系统可配置成:

获得用于要检索的所述记录的分片关键字,所述分片关键字基于所述记录的一个或多个字段;

针对与所述多个数据库分片中的每个数据库分片相关联的相应布隆过滤器中的每个布隆过滤器,通过将由所述布隆过滤器使用的一个或多个散列函数应用于所述分片关键字来生成与所述布隆过滤器一起使用的所述分片关键字的相应表示;

标识所述相应布隆过滤器中的哪些布隆过滤器与其的所述分片关键字的相应表示匹配;

搜索与所标识的布隆过滤器相关联的所述数据库分片以找到所述记录,其中所述记录被存储在基于计算的针对每个布隆过滤器的值和与所述布隆过滤器一起使用的所述分片关键字的所述相应表示之间的汉明距离而选择的数据库分片中;以及

提供所述记录。

用于数据库分片的方法和系统

背景技术

技术领域

[0001] 本发明涉及分片数据库。具体地,本发明涉及用于存储和检索分片数据库的分片表中的记录的方法、系统和计算机程序产品。

[0002] 相关技术的描述

[0003] 数据库分片是用于将数据库拆分成称为数据库分片的多个较小部分的已知技术。这种技术也可以被称为水平分割。

[0004] 通常,当将分片应用于数据库时,数据库分片中的每一个被配置为包括与所有其他数据库分片相同的模式(这可以是针对整个数据库或针对其某个子集的模式)。换言之,数据库分片中的每一个包括与其他数据库分片中的每一个相同的表定义。在应用数据库分片技术期间,某些表可以被配置为分片表。这意味着那些表的记录将被分布在数据库分片之间,使得存储在每个数据库分片上的那些表中的记录(或数据)对于每个数据库碎片将是不同的。通常基于从分片表的一个或多个字段导出的分片关键字的值来确定分片表的记录在数据库分片之间的分布。例如,分片关键字可以被定义为分片表的"ID"字段,并且具有落入第一范围内的"ID"的记录可以被存储在第一数据库分片上的表中,而具有落入不同的第二范围内的"ID"的记录可以被存储在第二数据库分片上的表中。因此,即使数据库分片可以具有彼此相同的模式,它们也仅包括模式内的任意分片表格的记录的子集。为了从分片表检索记录,首先需要确定哪个数据库分片存储感兴趣的记录。这可以例如通过确定分片关键字是落入第一范围还是第二范围以确定记录是否应当分别被存储在第一数据库分片上还是第二数据库分片上来实现。

[0005] 除了分片表格之外,分片数据库还可以包括复制表,其在每个数据库分片之间复制。这意味着数据库碎片中的每一个包括那些复制表的所有记录。通过将分片数据库的一些表包括为复制表,可以减少需要在不同数据库分片之间执行的查找的数量,从而提高分片数据库的性能。

[0006] 数据库分片可以用于提高数据库的性能。作为示例,在数据库存储非常大量的记录或接收非常大量的查询的情况下,数据库分片可以用来创建可以在单独的计算设备上操作的多个数据库分片,从而提高数据库系统存储大量记录或回答大量查询的能力。这是因为其上存储每个数据库分片的计算设备可以独立地处理关于存储在该数据库分片中的数据的子集的查询。

[0007] 为了最大化从数据库分片实现的益处,期望分片表格的记录在每个数据库分片之间均匀地分布,使得每个数据库分片所需的存储要求和计算能力大致平衡。此外,还期望使得附加数据库分片能够被动态地添加到分片数据库以允许满足计算能力或存储要求的任何增加。

[0008] 通常,用于将记录分配给分片数据库中的分片表的方法不导致记录被随机分布,使得每个数据库分片所需的存储要求和计算能力被平衡。例如,基于"ID"字段是落入第一

范围还是第二范围而在分片之间分布记录的方法意味着具有较低"ID"的那些记录可能被分组在相同的数据库碎片上。如果递增地分配"ID",使得较旧的记录具有较低的"ID",则如果较新的记录被更经常查询,则这可能导致数据库分片之间的计算负载的不平衡。

[0009] 另外,用于将记录分配给分片数据库中的分片表的典型方法或者不允许动态地添加附加数据库分片,或者在添加数据库分片时导致增加的复杂性,从而导致在从分片数据库检索记录时降低的性能。

发明内容

[0010] 根据本发明的一个方面,提供了一种用于将记录存储在分片数据库的分片表中的计算机实现的方法。分片数据库包括多个数据库分片。数据库分片各自被配置为存储分片表的记录的子集。数据库分片还各自与相应的布隆(Bloom)过滤器相关联。布隆过滤器被配置成提供关于与特定分片关键字相关联的记录是否可能已经被存储在与其相关联的数据库分片中的指示。该方法包括获得用于要被存储的记录的分片关键字。分片关键字基于记录的一个或多个字段。该方法还包括生成分片关键字的相应表示以供与多个相应布隆过滤器中的每一个一起使用。给定布隆过滤器的表示是通过将该布隆过滤器所使用的一个或多个散列函数应用于分片关键字来生成的。该方法还包括计算多个布隆过滤器中的每一个布隆过滤器的当前值和与该布隆过滤器一起使用的分片关键字的相应表示之间的汉明距离。所述方法进一步包括基于所述所计算汉明距离选择数据库分片且将所述记录存储在所述所选择数据库分片中。

[0011] 布隆过滤器是一种已知的数据结构,其可被用于测试元素是否是集合的成员。布隆过滤器是表示固定长度比特向量(或数组)中的集合的成员资格的概率数据结构。最初,当布隆过滤器所表示的集合为空时,向量中的每一比特都为"0"。除了固定长度的比特向量之外,每个布隆过滤器具有与其相关联的预定数目的不同散列算法。这些散列算法每个都将任意给定元素映射到向量中的特定比特上。当元素被添加到由布隆过滤器表示的集合时,布隆过滤器将每个散列算法应用于每个添加的元素,并且将由散列算法将添加的元素映射到的比特向量的对应比特设置为"1"。

[0012] 为了测试特定元素是否是由布隆过滤器表示的集合的成员,将散列算法应用于该元素,并且检查该元素被散列算法映射到的比特向量的比特。如果该元素被映射到的任意比特是"0",则该元素不是布隆过滤器所表示的集合的成员。然而,如果元素被映射到的比特都是'1',则该元素可以是布隆过滤器所表示的集合的成员。

[0013] 虽然布隆过滤器可用于明确地确定特定元素不是该集合的成员,但是它可能返回错误地指示特定元素是该集合的成员而实际上不是该集合的成员的假阳性。这是因为元素映射到的比特可能已经通过将其他元素插入到也映射到那些比特的集合中而被设置为'1'。换句话说,布隆过滤器可以为给定元素提供两个结果之一,即:(1)所述元素不是由所述布隆过滤器表示的集合的成员;或者(2)该元素可以是布隆过滤器所表示的集合的成员。由布隆过滤器所使用的比特向量中的比特数量和它所使用的不同散列算法的数量来确定布隆过滤器提供假阳性的可能性。因此,当实现布隆过滤器时,可以选择这些参数,以提供从布隆过滤器接收假阳性的期望可能性。

[0014] 由于布隆过滤器表示集合的成员资格的方式,不可能从布隆过滤器所表示的集合

中删除项目。然而,布隆过滤器的变型是已知的,其使得能够从所表示的集合中删除项目。特别地,计数(Counting)过滤器是一种布隆过滤器,其将向量中的每个比特扩展为替代地为n比特计数器,允许记录每个比特已经被插入的元素映射的次数。这使得能够通过确定向量上要删除的元素由散列算法映射到的位置并且递减在那些位置处的n比特计数器,来从由计数过滤器表示的集合中删除元素。

[0015] 汉明距离是两个等长字符串之间的差的度量。汉明距离度量将一个字符串改变为另一个字符串所需的替换的最小数量(即,两个字符串彼此不同的位置的数量)。例如,4比特二进制串1001和1010之间的汉明距离是2,因为最后两比特彼此不同而前两比特相同。

[0016] 发明人已经认识到,可以在分片数据库中与汉明距离一起使用布隆过滤器以改进分片数据库中的记录的分布。具体地,通过基于与每个数据库分片相关联的布隆过滤器与用于该过滤器的分片关键字的表示之间的汉明距离选择对其中存储记录的数据库分片,所提出的实施例可以在分片数据库的数据库分片之间更均匀地分布记录。这可以改进分片数据库的性能。

[0017] 此外,基于这样的汉明距离在数据库分片中分布(或分配)记录可以使得从数据库分片检索记录也能够被改进。例如,通过使用汉明距离来将记录分配给数据库分片,相似的数据更有可能被共同定位。布隆过滤器然后可以提供消除未存储特定记录的数据库分片的非常快速的方式。

[0018] 与数据库分片相关联的相应布隆过滤器中的每一个可以可选地使用相同数量的比特和相同的散列函数。通过使布隆过滤器在它们使用的比特的数量和散列函数方面相同,可能仅需要生成分片关键字的单个表示。该表示可以适于与布隆过滤器中的每一个一起使用。

[0019] 各个布隆过滤器中的每一个可以可选地是计数过滤器。计数过滤器的使用可以使得记录能够被删除并且使得它们的删除能够由计数过滤器来表示,从而防止由于过滤器匹配分片关键字的表示而针对删除的记录提供假阳性。

[0020] 所选择的数据库分片可以是与所计算的汉明距离是多个所计算的汉明距离之中的最小值的相应布隆过滤器相关联的数据库分片。通过将记录存储在数据库分片中,针对该数据库分片的布隆过滤器与该布隆过滤器所使用的记录的表示相距最小汉明距离,相信每个布隆过滤器可保持彼此更不同。这平均起来可以导致指示特定记录的匹配的较少的布隆过滤器,从而导致需要被搜索以检索记录的数据库分片的数量的减少。该技术还可用于平衡记录到每个数据库分片的分配,从而防止它们偏离太多。

[0021] 所提出的用于在数据库分片之间分布记录的方法还可以减少与每个数据库分片相关联的布隆过滤器的假阳性率。这是因为在其与布隆过滤器一起使用的表示方面最相似的记录更可能被存储在相同的数据库分片上,并且因此更可能由分片数据库中的相同布隆过滤器来表示。假阳性率的降低可以平均地减少需要被搜索以检索特定记录的数据库分片的数量。

[0022] 对与所计算的汉明距离是多个所计算的汉明距离之中的最小值的相应布隆过滤器相关联的数据库分片的选择,可以包括从与所计算的汉明距离是最小值的相应布隆过滤器相关联的那些数据库分片中随机选择数据库分片。这可以提供当多个布隆过滤器具有到其分片关键字的相应表示的相同最小汉明距离时选择数据库分片的有效方式。

[0023] 分片数据库可以可选地还包括每个数据库分片是否可用于存储新记录的指示。在这种情况下,为其生成分片关键字的相应表示的多个布隆过滤器可以不包括与被指示为不可用的数据库分片相关联的任何布隆过滤器。这意味着,变得太满或过载的数据库分片可被排除在存储任何附加记录的考虑之外,从而帮助将每个数据库分片的性能维持在最小水平以上。

[0024] 实施例可以可选地生成分片关键字的表示以供与关联于被指示为可用的所有数据库分片的相应布隆过滤器一起使用。以此方式,所提出的实施例可以确保记录跨可用数据库分片的更好分布。

[0025] 分片数据库可以可选地包括用于对已经插入到数据库分片中的每个数据库分片中的记录的相应数目进行计数的相应插入计数器。所提出的实施例可以针对被选择用于存储记录的数据库分片递增插入计数器。因此,该插入计数器可以提供已经被插入到数据库分片中的每一个中的记录的数目的计数。

[0026] 所提出的实施例可以可选地进一步包括确定已经插入到所选择的数据库分片中的记录的数量是否超过预定阈值,并且如果是,则可以将所选择的数据库分片标记为不可用于存储新记录。以此方式,各实施例可以防止数据库分片以可能降低分片数据库的性能的方式被过度利用。

[0027] 实施例可以可选地进一步包括将新数据库分片和相关联的布隆过滤器添加到分片数据库。用于将记录存储在分片数据库中的所提出的概念因此可以提供用于将新数据库分片插入到分片数据库中的灵活性。通过将新数据库分片动态地插入分片数据库中,可以增加分片数据库的容量和性能以满足需求。此外,使用汉明距离来选择要在其上存储插入的记录的数据库分片可以使得能够跨数据库分片分配记录以自动地重新平衡以并入新插入的数据库分片。

[0028] 根据本发明的另一方面,提供了一种用于从分片数据库的分片表中检索记录的计算机实现的方法。分片数据库包括多个数据库分片。数据库分片各自被配置为存储分片表的记录的子集。数据库分片还各自与相应的布隆过滤器相关联。布隆过滤器被配置成提供关于与特定分片关键字相关联的记录是否可能已经被存储在与其相关联的数据库分片中的指示。该方法包括获得用于要检索的记录的分片关键字。分片关键字基于记录的一个或多个字段。所述方法还包括生成所述分片关键字的相应表示以供与和所述多个数据库分片相关联的所述布隆过滤器中的每一个一起使用。给定布隆过滤器的表示是通过将该布隆过滤器所使用的一个或多个散列函数应用于分片关键字来生成的。该方法还包括标识相应布隆过滤器中的哪些布隆过滤器与它们相应的分片键的表示匹配。该方法还包括搜索与那些所标识的布隆过滤器相关联的数据库分片以找到并提供记录。

[0029] 通过使用与数据库分片中的每一个相关联的布隆过滤器,很可能在需要进行对剩余数据库分片的深入搜索之前,可以非常快速且高效地从考虑中排除大量的数据库分片。这可以提高从分片数据库的分片表中检索记录的效率。

[0030] 与数据库碎片相关联的相应布隆过滤器中的每一个可以可选地使用相同数量的比特和相同的散列函数。通过使布隆过滤器在它们使用的比特的数量和散列函数方面相同,可能仅需要生成分片关键字的单个表示。这样的表示然后可以适于与布隆过滤器中的每一个一起使用。

[0031] 根据本发明的另一方面,提供了一种用于存储分片数据库的系统。分片数据库包括多个数据库分片。数据库分片各自被配置为存储分片表的记录的子集。数据库分片还各自与相应的布隆过滤器相关联。布隆过滤器被配置成提供关于与特定分片关键字相关联的记录是否可能已经被存储在与其相关联的数据库分片中的指示。该系统包括多个计算系统。每个计算系统被配置成存储所述多个数据库分片中的一个或多个。该系统还包括控制器,该控制器被配置为根据上述用于从分片数据库的分片表检索记录的方法从分片数据库的分片表检索记录。

[0032] 控制器可以可选地被配置成根据用于将记录存储在分片数据库的分片表中的上述方法将记录存储在分片数据库的分片表中。

[0033] 分片数据库可以可选地还包括相应删除计数器。删除计数器可以维护已经从其相应的数据库分片删除的记录的数量的计数。可选地,控制器可以进一步被配置为确定已经从特定数据库分片删除的记录的数量是否超过该数据库分片的相应预定阈值。如果已删除的记录的数目超过相应的预定阈值,则控制器可基于当前存储在该数据库碎片中的记录来重建该数据库分片的布隆过滤器。通过一旦删除的数目超过预定阈值就重建用于数据库分片的布隆过滤器,控制器可以减少可能由基于已经被删除的记录指示匹配的布隆过滤器导致的假阳性的数目。

[0034] 根据本发明的另一方面,提供了一种用于将记录存储在分片数据库的分片表中的计算机程序产品。计算机程序产品包括计算机可读存储介质。计算机可读存储介质具有存储在其上的可由处理器执行的程序指令。程序指令可由处理器执行以使处理器执行用于将记录存储在分片数据库的分片表中的上述方法的步骤。

[0035] 根据本发明的另一方面,提供了一种用于从分片数据库的分片表中检索记录的计算机程序产品。计算机程序产品包括计算机可读存储介质。计算机可读存储介质具有存储在其上的可由处理器执行的程序指令。程序指令可由处理器执行以使处理器执行用于从分片数据库的分片表中检索记录的上述方法的步骤。

[0036] 根据另一方面,提供了一种用于将记录存储在分片数据库的分片表中的系统,所述分片数据库包括多个数据库分片,所述数据库分片中的每个数据库分片被配置为存储所述分片表的记录的子集并且与相应的布隆过滤器相关联,所述布隆过滤器被配置为提供关于与特定分片关键字相关联的记录是否可能已经被存储在该数据库分片中的指示,所述系统可配置成:获得用于要被存储的所述记录的分片关键字,所述分片关键字基于所述记录的一个或多个字段;针对多个相应布隆过滤器中的每一个布隆过滤器,通过将该布隆过滤器所使用的一个或多个散列函数应用于所述分片关键字来生成与该布隆过滤器一起使用的所述分片关键字的相应表示;计算所述多个布隆过滤器中的每一个布隆过滤器的当前值和与该布隆过滤器一起使用的所述分片关键字的相应表示之间的汉明距离;基于所计算的汉明距离来选择数据库分片;以及将所述记录存储在所选择的数据库分片中。

[0037] 根据另一方面,提供了一种用于从分片数据库的分片表检索记录的系统,所述分片数据库包括多个数据库分片,所述数据库分片中的每个数据库分片被配置为存储所述分片表的记录的子集并且与相应的布隆过滤器相关联,所述布隆过滤器被配置为提供关于与特定分片关键字相关联的记录是否可能已经被存储在该数据库分片中的指示,所述系统可配置成:获得用于要检索的所述记录的分片关键字,所述分片关键字基于所述记录的一个

或多个字段;针对与所述数据库分片中的每个数据库分片相关联的相应布隆过滤器中的每个布隆过滤器,通过将由该布隆过滤器使用的一个或多个散列函数应用于所述分片关键字来生成与该布隆过滤器一起使用的所述分片关键字的相应表示;标识所述相应布隆过滤器中的哪些布隆过滤器与其的所述分片关键字的相应表示匹配;搜索与所标识的布隆过滤器相关联的数据库分片以找到所述记录;以及提供记录。

[0038] 附图简要说明

[0039] 现在将参考附图仅通过示例的方式描述本发明的实施例,其中:

[0040] 图1示意性地示出了本发明的实施例可以在其上运行的示例性计算机系统;

[0041] 图2示意性地示出了根据本发明的实施例的示例性分片数据库;

[0042] 图3是示意性地示出根据本发明的实施例的用于将记录存储在分片数据库的分片表中的计算机实现的方法的图;以及

[0043] 图4是示意性地示出根据本发明的实施例的用于从分片数据库的分片表中检索记录的计算机实现的方法的图。

具体实施方式

[0044] 在以下描述和附图中,描述了本发明的某些实施例。然而,应当理解,本发明不限于所描述的实施例,并且一些实施例可以不包括下面描述的所有特征。然而,很明显,在不背离所附权利要求中提出的本发明的更宽的精神和范围的情况下,可以在此进行各种修改和改变。此外,在本申请的上下文中,在本发明的实施例构成方法的情况下,应当理解,这样的方法是用于由计算机执行的过程(即,是计算机可实现的方法)。因此,该方法的各个步骤反映了计算机程序的各个部分(例如,一个或多个算法的各个部分)。

[0045] 附图中的图1示意性地示出了本发明的实施例可以在其上运行的示例性计算机系统100。示例性计算机系统100包括计算机可读存储介质102、存储器104、处理器106和一个或多个接口108,它们都通过一条或多条通信总线110链接在一起。示例性计算机系统100可以采取常规计算机系统的形式,诸如例如台式计算机、个人计算机、膝上型计算机、平板计算机、智能电话、智能手表、虚拟现实头戴式设备、服务器、大型计算机等。

[0046] 计算机可读存储介质102和/或存储器104可以存储一个或多个计算机程序(或软件或代码)和/或数据。存储在计算机可读存储介质102中的计算机程序可以包括用于处理器106执行以便使计算机系统100工作的操作系统。存储在计算机可读存储介质102和/或存储器104中的计算机程序可以包括根据本发明的实施例的计算机程序,或者当由处理器106执行时使得处理器106执行根据本发明的实施例的方法的计算机程序。

[0047] 处理器106可以是适于执行一个或多个计算机可读程序指令的任意数据处理单元,所述计算机可读程序指令诸如属于存储在计算机可读存储介质102和/或存储器104中的计算机程序的那些。作为一个或多个计算机可读程序指令的执行的一部分,处理器106可以将数据存储到计算机可读存储介质102和/或存储器104和/或从计算机可读存储介质102和/或存储器104读取数据。处理器106可以包括单个数据处理单元或并行或彼此协作运行的多个数据处理单元。作为一个或多个计算机可读程序指令的执行的一部分,处理器106可以将数据存储到计算机可读存储介质102和/或存储器104和/或从计算机可读存储介质102和/或存储器104读取数据。

[0048] 一个或多个接口108可以包括使计算机系统100能够通过网络与其它计算机系统通信的网络接口。网络可以是适于从一个计算机系统向另一个计算机系统发送或传送数据的任何种类的网络。例如,网络可以包括局域网、广域网、城域网、因特网、无线通信网络等中的一个或多个。计算机系统100可以经由任意适当的通信机制/协议通过网络与其它计算机系统通信。处理器106可以经由一条或多条通信总线110与网络接口通信,以使得网络接口通过网络向另一计算机系统发送数据和/或命令。类似地,一个或多个通信总线110使得处理器106能够对由计算机系统100经由网络接口从网络上的其它计算机系统接收的数据和/或命令进行操作。

[0049] 接口108可以替代地或附加地包括用户输入接口和/或用户输出接口。用户输入接口可以被布置为从系统100的用户或操作者接收输入。用户可以经由一个或多个用户输入设备(未示出,例如鼠标或其他定点设备、跟踪球或键盘)提供该输入。用户输出接口可以被布置成在显示器、监视器或屏幕(未示出)上向系统100的用户或操作者提供图形/视觉输出。处理器106可以指示用户输出接口形成图像/视频信号,该图像/视频信号使得显示器示出期望的图形输出。显示器可以是触敏的,使用户能够通过触摸或按压显示器来提供输入。

[0050] 应当理解,图1所示和以上所述的计算机系统100的体系结构仅仅是示例性的,并且可以使用具有使用替换组件或使用更多或更少组件的不同体系结构的系统来代替。

[0051] 附图中的图2示意性地示出了根据本发明的实施例的示例性分片数据库200。分片数据库200包括多个数据库分片210。例如,图2中所示的分片数据库200包括第一数据库分片210(1)和第二数据库分片210(2)。然而,将理解,也可使用额外的数据库分片210。数据库分片210中的每一个包括属于分片数据库200的模式的一个或多个表220的公共集合。将理解,如果需要,各个数据库分片也可包括额外的表(例如,作为数据库的单独垂直分区的一部分)。

[0052] 一个或多个表220的公共集合包括分片表220(1)。虽然分片表220(1)的模式在多个数据库分片210之间是相同的,但是存储在每个数据库分片210上的分片表220(1)的副本中的记录230的集合是不同的。具体地,数据库分片210中的每一个存储属于分片表220(1)的记录的子集。例如,图2中所示的分片数据库200具有属于存储在第一数据库分片210(1)上的分片表220(1)的第一记录子集230(1)和存储在第二数据库分片210(2)上的第二记录子集230(2)。然而,将理解,如果分片数据库200使用更多数据库分片210,则分片表220的记录集合230可以被分解成更多子集并且相应地跨数据库分片210分布。

[0053] 属于表220的公共集合的一些表220(2)和220(3)可以是复制的表。这意味着每个数据库分片210上的复制表220(2)和220(3)的副本是相同的,并且包括彼此相同的记录集。然而,将理解,表的公共集合220不必一定包括任何复制的表。

[0054] 分片数据库200还包括一个或多个布隆过滤器240,多个数据库分片210中的每一个有一个布隆过滤器。布隆过滤器240中的每一个被配置成提供关于共享表220(1)的特定记录是否已被存储在与其相关联的数据库分片210中的指示。为此,布隆过滤器240维护表示已被插入到其相应数据库分片210中的记录的子集230(1)或230(2)的固定大小的比特向量。例如,图2中所示的分片数据库200具有与第一数据库分片210(1)相关联的第一布隆过滤器240(1)和与第二数据库分片210(2)相关联的第二布隆过滤器240(2),每个布隆过滤器240的比特向量包括由于已经存储在其相应数据库分片210中的记录而已经被设置为'1'的

一些比特(由图2中所示的比特向量的表格表示上的黑色阴影表示)。将理解,在分片数据库200包括额外的数据库分片210的情况下,每个额外的数据库分片210将与其自己的表示存储在该数据库分片210中的记录的子集的布隆过滤器240相关联。如下文进一步讨论的,由布隆过滤器中的每一个维护的记录的子集230(1)和230(2)的表示基于每个记录的分片关键字260。

[0055] 现在将参考图3进一步讨论图2中所示的分片数据库200,该图是示意性地示出根据本发明的实施例的用于将记录250存储在分片数据库200的分片表220(1)中的计算机实现的方法300的图。

[0056] 在步骤310,方法300获得要存储的记录250的分片关键字260。分片关键字260基于被插入分片表格220(1)中的记录的一个或多个字段。在其最简单的形式中,分片关键字260可以从诸如"ID"字段之类的单个字段中导出。然而,可以替代地使用其他更复杂的复合分片关键字260,包括例如从每个记录的一些或所有字段导出的复合关键字。分片关键字260可以与要存储的记录一起提供,或者可以从要存储的记录的一个或多个字段生成。分片关键字260可以用于属于分片表220(1)的记录集合230中唯一地标识分片表220(1)的每个记录。然而,情况不必如此。相反,可以形成分片关键字,使得多个不同的记录可以与相同的分片关键字相关联。

[0057] 在步骤320处,方法300生成与数据库分片210相关联的布隆过滤器240一起使用的分片关键字260的一个或多个表示270。分片关键字260的这个表示270是通过将每个布隆过滤器240所使用的散列函数应用于分片关键字260而生成的,从而导致生成分片关键字260的比特向量表示270,其长度与由每个布隆过滤器240维护的记录子集230(1)和230(2)的比特向量表示的长度相同。用于每个数据库分片210的布隆过滤器240可以在比特向量的长度和它们使用的散列函数方面彼此相同(尽管每个布隆过滤器240的值当然将取决于已经插入每个相关联的数据库分片210中的记录而不同)。在这种情况下,仅需要生成分片关键字260的单个表示270,因为表示270对于所有布隆过滤器240将是相同的。然而,将理解,布隆过滤器240中的一些(或全部)可在结构上不同,具有不同长度的比特向量或利用不同(或不同数量的)散列算法。在这种情况下,将需要为布隆过滤器240的每个不同构造生成分片关键字260的表示270。在图2中所示的示例性分片数据库200中,例如,可以生成用于要存储的记录250的分片关键字260的单个表示270以供两个布隆过滤器240使用,因为它们是相同的。在该示例中,将用于布隆过滤器240的散列算法应用于要插入的记录250的分片关键字260生成了比特向量表示'100100000'。然而,将理解,不同的记录将导致生成不同的比特向量表示。

[0058] 在步骤330,方法300计算为每个布隆过滤器240生成的分片关键字260的比特向量表示270和布隆过滤器240(表示已经存储在每个数据库分片210上的记录的子集230(1)和230(2))的值之间的汉明距离。返回到图2中所示的示例性分片数据库200,存储在由第一布隆过滤器240(1)维护的第一数据库分片210(1)中的项目的第一子集230(1)的比特向量表示是'101000100',而存储在由第二布隆过滤器240(2)维护的第二数据库分片210(2)中的项目的第二子集230(2)的比特向量表示是'000001010'。因此,在该示例中,要存储的记录250的分片关键字260的表示270与第一布隆过滤器240(1)之间的汉明距离是"3"(因为它们在3个单独的地方不同),而到第二布隆过滤器240(2)的汉明距离是"4"(因为它们在4个地

方不同)。

[0059] 在步骤340处,方法300基于所计算的汉明距离选择数据库分片210。也就是说,数据库分片210的选择是基于汉明距离的函数。作为示例,对于被存储的记录250,所选择的数据库分片210可以是与具有距分片关键字260的表示270的最小汉明距离的布隆过滤器240相关联的数据库分片210。将理解,在一些情况下,仅单个数据库分片210将与具有最小汉明距离的布隆过滤器相关联。在这样的情况下,数据库分片210的选择可以是自动的(即,与具有最小汉明距离的布隆过滤器相关联的单个数据库分片210)。然而,在其它情况下,最小汉明距离可由多个布隆过滤器240提供。例如,若干布隆过滤器可具有相同的汉明距离,该汉明距离是所计算的汉明距离中的最小距离。在这样的情况下,数据库分片210可以从其布隆过滤器具有最小汉明距离准则的那些数据库分片中选择。例如,数据库分片210可以从具有相同汉明距离的那些数据库分片中随机选择,所述相同汉明距离是所计算的汉明距离中的最小距离。然而,可以替代地使用用于选择数据库分片210的其他手段。例如,可以从满足最小汉明距离准则的那些数据库分片210之中选择其上存储了最少记录的数据库分片210。基于汉明距离进行选择的各种其它因素或方法对于本领域技术人员来说是显而易见的。返回图2中所示的示例,第一数据库分片210(1)被选择作为用于存储记录250的位置,因为分片关键字260的表示270与第一布隆过滤器240(1)之间的汉明距离具有针对第一和第二布隆过滤器240(1)和240(2)分别计算的汉明距离'3'和'4'中的最小值'3'。

[0060] 在步骤350处,方法300将记录250存储在所选择的数据库分片210中。当记录250被存储在数据库分片210中时,与数据库分片210相关联的布隆过滤器240被更新以表示该记录250现在是由该布隆过滤器240表示的项目的子集的一部分。在图2中所示的示例中,第一布隆过滤器240(1)因此将被更新以添加记录250的分片关键字260的比特向量表示270,使得由布隆过滤器240(1)维护的比特向量表示变为'101100100'。通过每当记录被存储在与它们相关联的数据库分片210中时更新布隆过滤器240,它们所维护的记录子集的表示将准确地反映存储在每个数据库分片210上的记录。

[0061] 随着更多记录被存储在分片数据库200的数据库分片210上,数据库分片210中的一些可以达到其中期望防止任何另外的记录被存储在那些数据库分片210上以防止那些数据库分片210(并且因此分片数据库200作为整体)的性能降到某一水平以下的水平。为了对此进行辅助,分片数据库200还可以包括每个数据库分片是否可用于存储新记录或不可用于存储新记录的指示(或标记)。例如,分片数据库200可以包括与每个数据库分片210相关联的二进制标志以表示数据库分片210是否可用(尽管可以替代地使用用于指示数据库分片210对于存储新记录的可用性的其他技术)。当存在用于存储新记录的每个数据库分片210的可用性的这种指示时,方法300可不考虑被指示为不可用于存储新记录的那些数据库分片210。也就是说,当方法300在步骤320生成分片关键字260的一个或多个表示270时,它可以不生成供与被指示为不可用的数据库分片210相关联的任何布隆过滤器240使用的表示。相反,方法300可仅生成表示270以供被指示为可用于存储新记录的布隆过滤器240中的每一个(或其某个子集)一起使用。方法300然后可继续进行以计算已针对其生成表示270的布隆过滤器中的每一个的汉明距离,使得在基于汉明距离选择数据库分片时将仅考虑可用于存储新记录的那些数据库分片210。

[0062] 为了确定存储在数据库分片210中的记录的数目何时已经达到其应当被标记为不

可用于存储新记录的水平,分片数据库200可以包括与数据库分片210中的每个数据库分片相关联的插入计数器。通过每当记录被存储在其所关联的数据库分片210上时递增插入计数器,插入计数器可以被用于确定在任何给定时间已经被插入到数据库分片210中的每一个中的记录的当前数目。可以为数据库分片210中的每一个可以存储的记录的数量设置预定阈值。该预定阈值对于分片数据库200中的所有数据库分片210可以是相同的,或者可以针对每个数据库分片210被不同地设置以考虑该数据库分片210可用的资源的任何差异。方法300因此可以确定已被插入到数据库分片210中的任一个中的记录的数量是否超过该数据库分片210的预定阈值,并且如果是,则可以将数据库分片210标记为不可用于存储任何新记录。每当方法300存储记录时,通过考虑在步骤340处选择的数据库分片210中存储的记录的数量是否超过该数据库分片210的预定阈值,可以执行该确定。然而,该确定还可以独立于用于存储新记录的方法300来执行,诸如例如由负责在周期性基础上管理分片数据库200的数据库管理系统来执行。

[0063] 为了增加分片数据库200的容量或性能,可以将额外的数据库分片210动态地添加到分片数据库200。例如,这可以作为方法300的一部分在数据库分片210被标记为不可用于存储新记录的任何时候执行,以便维持可用于存储新记录的预定数量的数据库分片210。然而,新数据库分片210的插入也可以独立于方法300而执行,诸如由负责管理分片数据库200的数据库管理系统来执行。这样的数据库管理系统可以例如连续地监控分片数据库200并且根据需要向分片数据库200添加新的数据库分片210以便维持预定的性能水平。

[0064] 当将新数据库分片210添加到分片数据库200时,还添加与新数据库分片210相关联的新的空布隆过滤器240。由于分片数据库200中的布隆过滤器240的操作,新记录将开始以用于跨可用于存储新记录的那些数据库分片210来平衡存储的记录的方式被存储在新数据库分片210中。这是因为,其上存储有更多记录的数据库分片210的布隆过滤器可能包括被设置为"1"的大量比特。因此,存储在先前存在的数据库分片210上的记录越多,任何新记录的分片关键字260的表示270,就汉明距离而言,将越可能比先前存在的布隆过滤器240更接近新添加的数据库分片210的空布隆过滤器240(其最初使其比特向量的所有比特设置为'0')。这意味着,相比预先存在的数据库分片,最初更可能在新数据库分片上存储新记录。

[0065] 图4是示意性地示出用于从分片数据库200的分片表220(1)检索记录的计算机实现的方法400的图。

[0066] 在步骤410,方法400获得要被检索的记录的分片关键字260。如上文关于用于将记录存储在分片数据库200中的方法300的步骤310所述,分片关键字260基于记录的一个或多个字段。用于要检索的记录的分片关键字260可以作为方法400的输入来提供,例如,作为检索与该分片关键字260相关联的记录的请求280的一部分。或者,分片关键字260可以从作为关于要检索的(一个或多个)记录的请求280的一部分提供的数据中导出。

[0067] 在步骤420,方法400生成与布隆过滤器240一起使用的分片关键字260的表示270。如上文关于用于将记录存储在分片数据库200中的方法300的步骤320所述,通过将用于特定布隆过滤器240的散列函数应用于分片关键字260来生成用于该布隆过滤器240的分片关键字260的表示。由于在检索记录(包括被标记为对于存储新记录为非活动的任何记录)时将需要考虑数据库分片210中的每一个,因此分片关键字260的表示270被生成以供与布隆

过滤器中的每一个使用。然而,再次,在用于每个数据库分片210的布隆过滤器240彼此具有相同形状的情况下(即,在用于每个布隆过滤器的比特向量的长度和散列函数相同的情况下),则仅需要生成分片关键字260的单个表示270,因为表示270对于所有那些布隆过滤器240将是相同的。

[0068] 在步骤430,方法400识别相应布隆过滤器240中的哪些与其的分片关键字260的相应的表示270匹配。为了确定分片关键字260的表示270是否匹配其相应的布隆过滤器240,分析布隆过滤器240的比特向量以确定分片关键字260被散列算法映射到的位置处的比特中的任何比特是否为"0"。如果分片关键字260映射到其上的比特中的任何比特是'0',则分片关键字270的表示270不匹配布隆过滤器240,从而指示分片关键字270所涉及的记录尚未存储在与该布隆过滤器240相关联的数据库分片210上。然而,如果所有的比特都是"1",则分片关键字260的表示270匹配布隆过滤器240,指示分片关键字260所涉及的记录可以被存储在与该布隆过滤器240相关联的数据库分片210上。

[0069] 将理解,在方法400的步骤420处的表示270的生成可以作为在方法400的步骤430处确定与分片关键字260匹配的布隆过滤器的一部分而被隐式地执行。也就是说,方法400可通过将每个布隆过滤器的散列算法应用于分片关键字260来确定分片关键字260映射到每个布隆过滤器的比特向量的哪些位置,从而隐式地计算表示270,而不是显式地计算分片关键字260的比特向量表示270。

[0070] 在步骤440处,方法400搜索与所识别的布隆过滤器240相关联的数据库分片210以找到记录。由于布隆过滤器240的概率性质,可能的是,布隆过滤器240中的多于一个可以指示与分片关键字260的表示270的匹配。这在一个或多个布隆过滤器240提供假阳性时发生。因此,方法400搜索与指示与分片关键字260的表示270匹配的布隆过滤器240相关联的数据库分片210中的每一个以便找到记录。

[0071] 在步骤450,方法400提供记录。该记录可以直接响应于来自外部系统的请求而提供,或者可以作为对作为更大的查询操作的一部分的内部请求的响应而提供。

[0072] 为了更清楚地描述本发明,使用具有单个分片表格220(1)的分片数据库200,已经讨论并在图2中示出了上述分片数据库200以及相关联的方法300和400。然而,将理解,分片数据库200可以包括其记录分布在数据库分片210之间的多个分片表格220。在这种情况下,多个布隆过滤器240可与每个数据库分片210相关联,一个布隆过滤器240针对每个分片表220。用于存储和从分片表检索记录的方法300和400然后可以参考用于方法在其上操作的分片表220的相关布隆过滤器240。

[0073] 上述分片数据库200可以存储在包括多个计算机系统(或设备)100的系统上。分片数据库200的数据库分片210中的每个数据库分片可以被存储在系统的相应计算机系统100上,使得计算机系统100中的每个计算机系统存储多个数据库分片210中的一个或多个数据库分片。系统还包括控制器290,其被配置成执行上述用于从分片数据库200的分片表220(1)检索记录的方法400。如图2中所示,控制器290可以在逻辑上与数据库分片210分离,并且可以在分离的计算机系统100上实现。或者,控制器290可以在也用于存储数据库分片210中的一个或多个的单个计算机系统100上实现。然后,可以将从分片数据库200的分片表220(1)检索记录的任何请求280定向到单个控制器290。或者,控制器290可以分布在用于存储数据库分片210中的一个或多个的多个(或所有)计算机系统100之间。在这种情况下,从分

片数据库220的分片表220(1)检索记录的请求280可以被指向它所位于的计算机系统100中的任何一个上的控制器。

[0074] 控制器290还可以被配置成使用上述方法300将记录250存储在分片数据库200的分片表220(1)中。因此,存储记录250的请求也可以被指向控制器290驻留于其上的任何计算机系统100。

[0075] 控制器290还可以被配置成允许从分片数据库200的分片表220(1)删除记录。这可以通过定位记录所位于的数据库分片210并且从存储在该数据库分片210上的分片表220(1)的记录的子集230移除记录来执行。然而,由于布隆过滤器的限制,不可能更新与该数据库分片210相关联的布隆过滤器240以反映从其所表示的记录的集合中移除该记录。因此,从数据库分片210删除的记录越多,则该数据库分片210的布隆过滤器240在尝试从分片表220(1)检索记录时将返回假阳性的可能性越大。为了解决这个问题,分片数据库200可以包括与数据库分片210中的每个数据库分片相关联的删除计数器。通过每当记录从其所关联的数据库分片210中被删除时递增删除计数器,删除计数器可以用于确定在不再存储在其关联的数据库分片210上的记录集合的每个布隆过滤器的表示中包括多少记录。控制器290然后可以确定已经从特定数据库分片210删除的记录的数量是否超过该分片的预定阈值,并且如果是,则可以从当前存储在该数据库分片210中的记录重建布隆过滤器240。然后可以重置与重建布隆过滤器相关联的数据库分片210的删除计数器。

[0076] 重建布隆过滤器240基本上涉及重置其比特向量,使得每一比特被设置为'0',并且通过将布隆过滤器240的散列算法重新应用于所存储的记录的分片关键字260,并将布隆过滤器的适当比特设置为'1',将该数据库分片210中存储的分片表220(1)的记录中的每一个重新添加到比特向量。在重建布隆过滤器240的同时,数据库分片210可以被临时标记为不可用于存储新记录,从而即使在与数据库分片210中的一个或多个相关联的布隆过滤器240可以被重建的同时,允许分片数据库200继续接受新记录以供存储。类似地,在布隆过滤器240被重建之前,其比特向量(或数组)的当前值可以被高速缓存,从而允许分片数据库200继续从分片表220(1)检索记录,即使在与数据库分片210中的一个或多个相关联的布隆过滤器240可能正被重建时也是如此。

[0077] 在替代实施例中,上述分片数据库200可以包括计数过滤器而不是布隆过滤器。如上所述,计数过滤器允许从由计数过滤器维护的记录集的表示中移除记录。因此,在该替代实施例中,当记录从分片表220(1)被删除时,控制器290可以更新与记录被从其删除的数据库分片210相关联的计数过滤器,而不需要重建过滤器。为了计算到计数滤波器的汉明距离,计数滤波器的值可以被平坦化成比特向量表示。换句话说,在由计数滤波器维护的向量中的每个位置处的n比特计数器可以被转换成单个比特表示,如果n比特计数器的值是'0',则该单个比特表示是'0',或者如果n比特计数器的值大于'0',则该单个比特表示是'1'。作为示例,如果计数滤波器的当前值是"01023002",则为了计算汉明距离的目的,这可以被平坦化成为比特向量表示"01011001"。

[0078] 在任何可能的技术细节结合层面,本发明可以是系统、方法和/或计算机程序产品。计算机程序产品可以包括计算机可读存储介质,其上载有用于使处理器实现本发明的各个方面的计算机可读程序指令。

[0079] 计算机可读存储介质可以是可以保持和存储由指令执行设备使用的指令的有形

设备。计算机可读存储介质例如可以是——但不限于——电存储设备、磁存储设备、光存储设备、电磁存储设备、半导体存储设备或者上述的任意合适的组合。计算机可读存储介质的更具体的例子(非穷举的列表)包括:便携式计算机盘、硬盘、随机存取存储器(RAM)、只读存储器(ROM)、可擦式可编程只读存储器(EPROM或闪存)、静态随机存取存储器(SRAM)、便携式压缩盘只读存储器(CD-ROM)、数字多功能盘(DVD)、记忆棒、软盘、机械编码设备、例如其上存储有指令的打孔卡或凹槽内凸起结构、以及上述的任意合适的组合。这里所使用的计算机可读存储介质不被解释为瞬时信号本身,诸如无线电波或者其他自由传播的电磁波、通过波导或其他传输媒介传播的电磁波(例如,通过光纤电缆的光脉冲)、或者通过电线传输的电信号。

[0080] 这里所描述的计算机可读程序指令可以从计算机可读存储介质下载到各个计算/处理设备,或者通过网络、例如因特网、局域网、广域网和/或无线网下载到外部计算机或外部存储设备。网络可以包括铜传输电缆、光纤传输、无线传输、路由器、防火墙、交换机、网关计算机和/或边缘服务器。每个计算/处理设备中的网络适配卡或者网络接口从网络接收计算机可读程序指令,并转发该计算机可读程序指令,以供存储在各个计算/处理设备中的计算机可读存储介质中。

[0081] 用于执行本发明操作的计算机程序指令可以是汇编指令、指令集架构(ISA)指令、机器指令、机器相关指令、微代码、固件指令、状态设置数据、集成电路配置数据或者以一种或多种编程语言的任意组合编写的源代码或目标代码,所述编程语言包括面向对象的编程语言—诸如Smalltalk、C++等,以及过程式编程语言—诸如“C”语言或类似的编程语言。计算机可读程序指令可以完全地在用户计算机上执行、部分地在用户计算机上执行、作为一个独立的软件包执行、部分在用户计算机上部分在远程计算机上执行、或者完全在远程计算机或服务器上执行。在涉及远程计算机的情形中,远程计算机可以通过任意种类的网络—包括局域网(LAN)或广域网(WAN)—连接到用户计算机,或者,可以连接到外部计算机(例如利用因特网服务提供商来通过因特网连接)。在一些实施例中,通过利用计算机可读程序指令的状态信息来个性化定制电子电路,例如可编程逻辑电路、现场可编程门阵列(FPGA)或可编程逻辑阵列(PLA),该电子电路可以执行计算机可读程序指令,从而实现本发明的各个方面。

[0082] 这里参照根据本发明实施例的方法、装置(系统)和计算机程序产品的流程图和/或框图描述了本发明的各个方面。应当理解,流程图和/或框图的每个方框以及流程图和/或框图中各方框的组合,都可以由计算机可读程序指令实现。

[0083] 这些计算机可读程序指令可以提供给通用计算机、专用计算机或其它可编程数据处理装置的处理器,从而生产出一种机器,使得这些指令在通过计算机或其它可编程数据处理装置的处理器执行时,产生了实现流程图和/或框图中的一个或多个方框中规定的功能/动作的装置。也可以把这些计算机可读程序指令存储在计算机可读存储介质中,这些指令使得计算机、可编程数据处理装置和/或其他设备以特定方式工作,从而,存储有指令的计算机可读介质则包括一个制造品,其包括实现流程图和/或框图中的一个或多个方框中规定的功能/动作的各个方面的指令。

[0084] 也可以把计算机可读程序指令加载到计算机、其它可编程数据处理装置、或其它设备上,使得在计算机、其它可编程数据处理装置或其它设备上执行一系列操作步骤,以产

生计算机实现的过程,从而使得在计算机、其它可编程数据处理装置、或其它设备上执行的指令实现流程图和/或框图中的一个或多个方框中规定的功能/动作。

[0085] 附图中的流程图和框图显示了根据本发明的多个实施例的系统、方法和计算机程序产品的可能实现的体系架构、功能和操作。在这点上,流程图或框图中的每个方框可以代表一个模块、程序段或指令的一部分,所述模块、程序段或指令的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。在有些作为替换的实现中,方框中所标注的功能也可以以不同于附图中所标注的顺序发生。例如,两个连续的方框实际上可以基本并行地执行,它们有时也可以按相反的顺序执行,这依所涉及的功能而定。也要注意的是,框图和/或流程图中的每个方框、以及框图和/或流程图中的方框的组合,可以用执行规定的功能或动作的专用的基于硬件的系统来实现,或者可以用专用硬件与计算机指令的组合来实现。

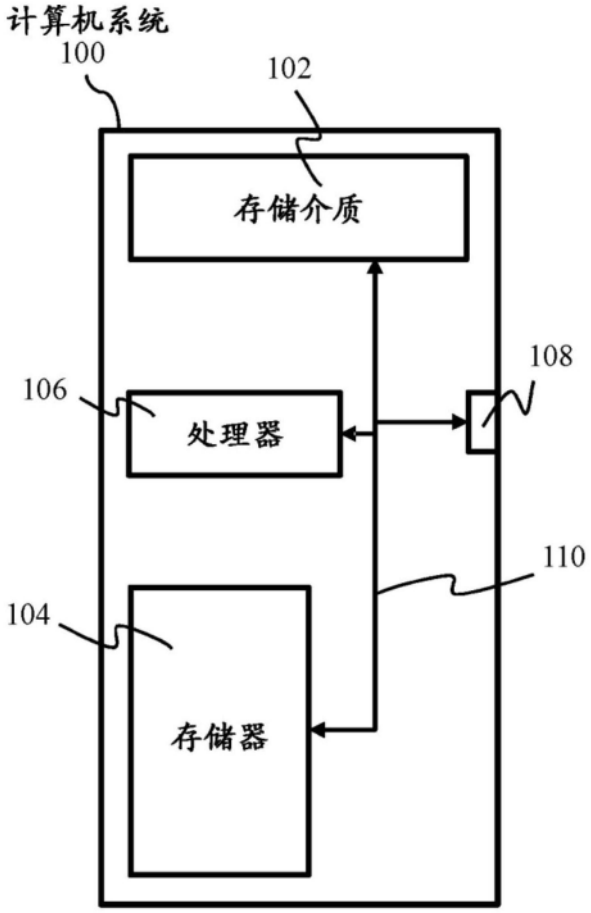


图1

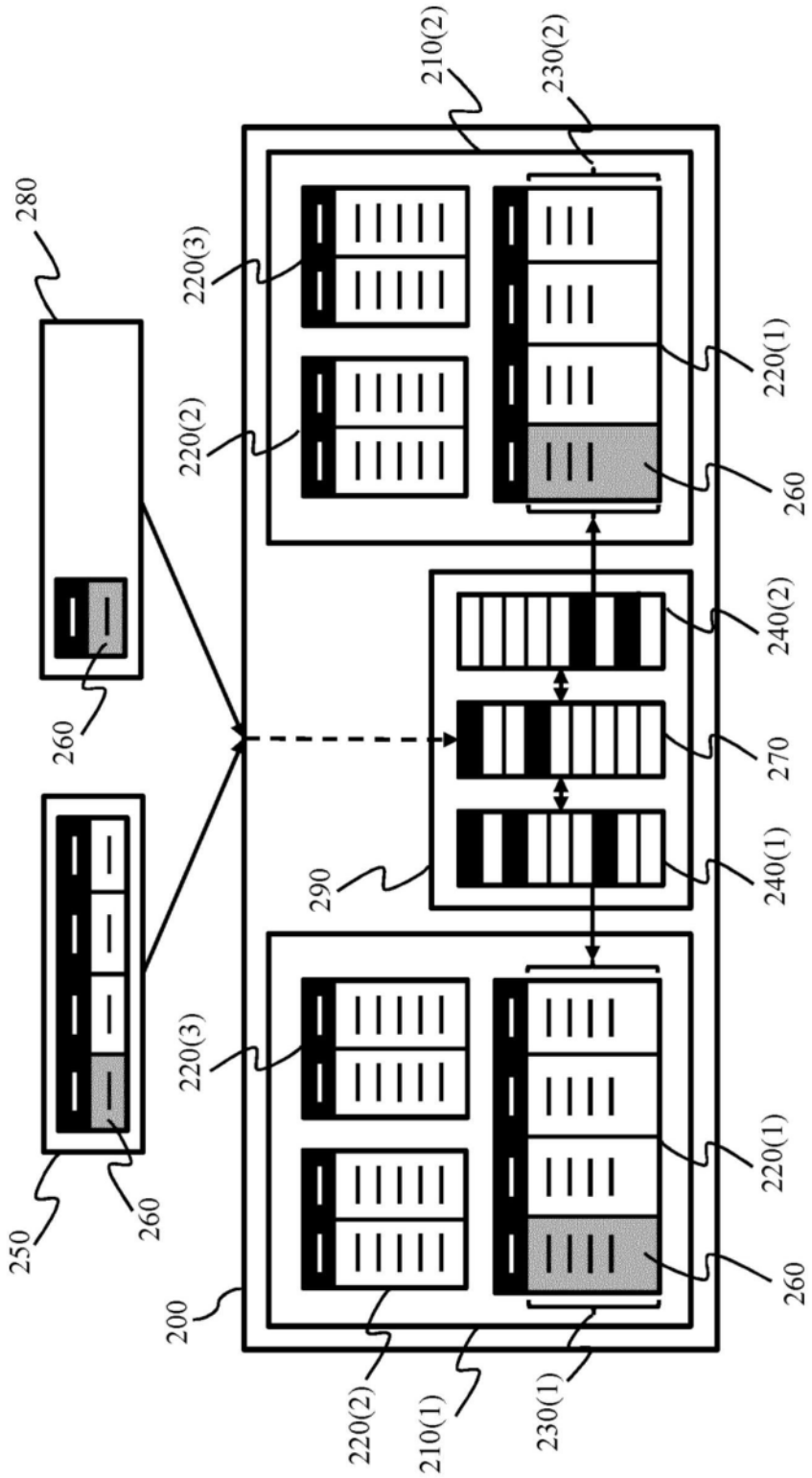

210(2)
230(2)
280
220(3)
220(1)
220(2)
260
240(2)
270
260
290
240(1)
220(3)
220(1)
250
260
260
200
220(2)
210(1)
230(1)

图2

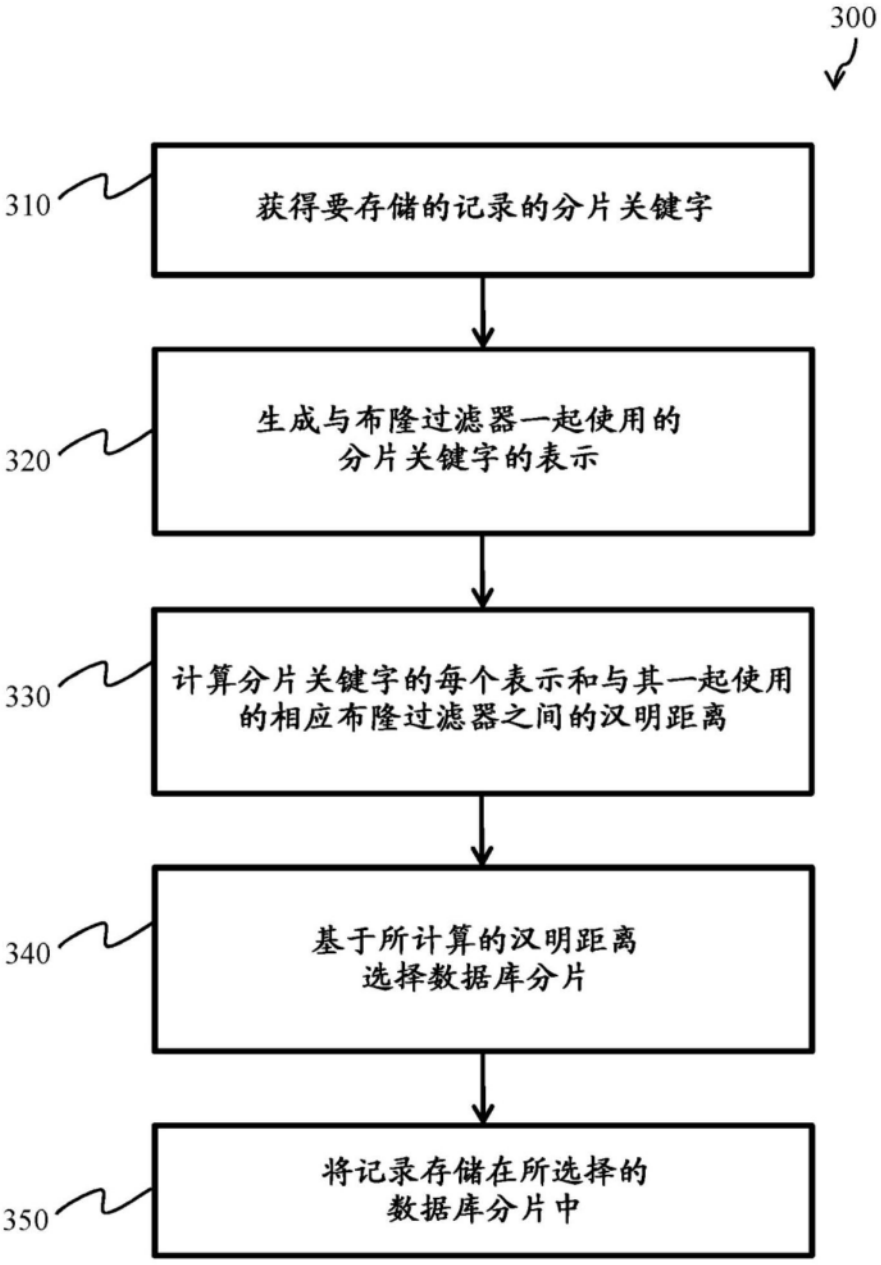

300
310
获得要存储的记录的分片关键字
320
生成与布隆过滤器一起使用的
分片关键字的表示
330
计算分片关键字的每个表示和与其一起使用
的相应布隆过滤器之间的汉明距离
340
基于所计算的汉明距离
选择数据库分片
350
将记录存储在所选择的
数据库分片中

图3

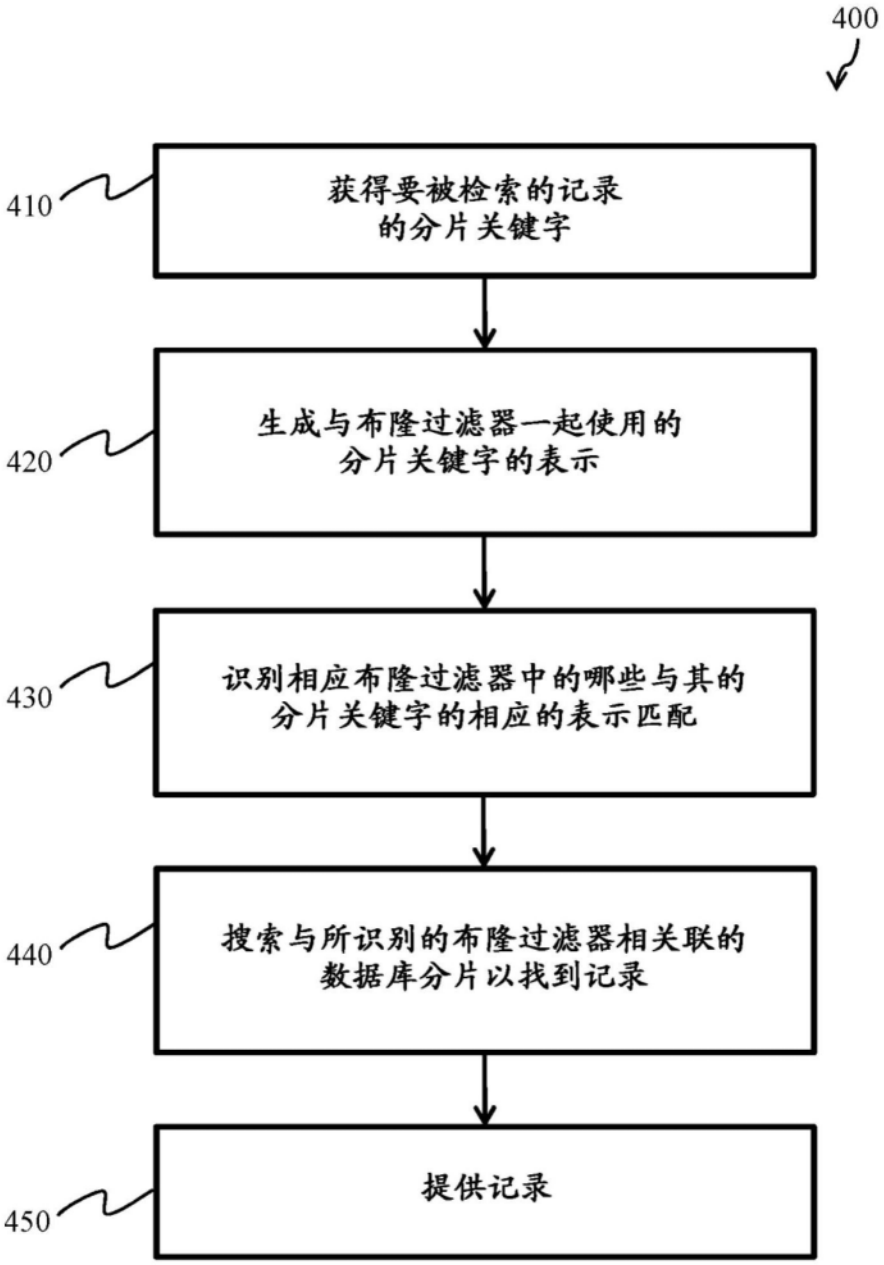

400
410
获得要被检索的记录的分片关键字
420
生成与布隆过滤器一起使用的分片关键字的表示
430
识别相应布隆过滤器中的哪些与其的分片关键字的相应的表示匹配
440
搜索与所识别的布隆过滤器相关联的数据库分片以找到记录
450
提供记录

图4