



(12) 发明专利申请

(10) 申请公布号 CN 106503010 A

(43) 申请公布日 2017. 03. 15

(21) 申请号 201510564356. 2

(22) 申请日 2015. 09. 07

(71) 申请人 北京国双科技有限公司

地址 100086 北京市海淀区双榆树小区知春
路 76 号翠宫饭店 8 层 A 间

(72) 发明人 石岱曦

(74) 专利代理机构 北京鼎佳达知识产权代理事

务所(普通合伙) 11348

代理人 王伟锋 刘铁生

(51) Int. Cl.

G06F 17/30(2006. 01)

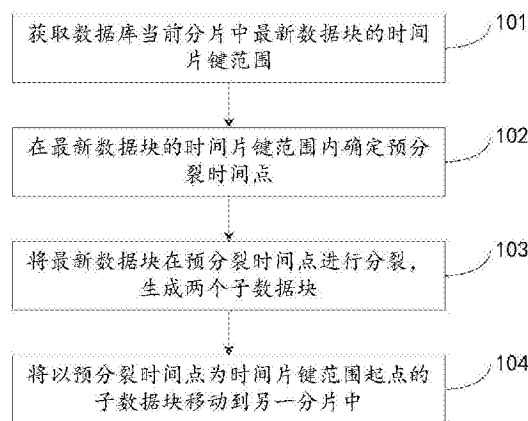
权利要求书2页 说明书7页 附图1页

(54) 发明名称

一种数据库更改写入分区的方法及装置

(57) 摘要

本发明公开了一种数据库更改写入分区的方法及装置,涉及数据处理领域,解决了在数据库分片集群中选取单调递增片键进行分区时导致数据分片不均衡的问题。本发明的方法包括:获取数据库当前分片中最新数据块的时间片键范围,最新数据块为当前写入数据的数据块,在该时间片键范围内确定预分裂时间点,将所述最新数据块在预分裂时间点进行分裂,生成两个子数据块,将以所述预分裂时间点为时间片键范围起点的子数据块移动到另一分片中。本发明主要用于数据库对写入的数据进行均衡分布。



1. 一种数据库更改写入分区的方法,其特征在于,所述方法包括:

获取数据库当前分片中最新数据块的时间片键范围,所述最新数据块为当前写入数据的数据块;

在所述最新数据块的时间片键范围内确定预分裂时间点;

将所述最新数据块在所述预分裂时间点进行分裂,生成两个子数据块;

将以所述预分裂时间点为时间片键范围起点的子数据块移动到另一分片中。

2. 根据权利要求 1 所述的方法,其特征在于,所述获取数据库当前分片中最新数据块的时间片键范围,包括:

监控所述数据库分片集群中当前分片的数据量,当所述当前分片的数据量大于预设阈值时,获取所述当前分片中最新数据块的时间片键范围。

3. 根据权利要求 1 所述的方法,其特征在于,在所述最新数据块的时间片键范围内确定预分裂时间点,包括:

在所述最新数据块的时间片键范围内,将当前时间点之后的任一个时间点确定为预分裂时间点。

4. 根据权利要求 1 所述的方法,其特征在于,将所述最新数据块在所述预分裂时间点进行分裂,包括:

基于分裂操作指令的触发将所述最新数据块在所述预分裂时间点进行分裂。

5. 根据权利要求 1 所述的方法,其特征在于,所述生成两个子数据块,包括:

生成子数据块 A 和子数据块 B;

所述子数据块 A 以分裂前的所述最新数据块的时间片键范围的起点和所述预分裂时间点作为自身时间片键范围的起点和终点;

所述子数据块 B 以所述预分裂时间点和分裂前的所述最新数据块的时间片键范围的终点作为自身时间片键范围的起点和终点;

将以所述预分裂时间点为时间片键范围起点的子数据块移动到另一分片中,包括:

将以所述预分裂时间点为时间片键范围起点的子数据块 B 移动到另一分片中。

6. 根据权利要求 1 所述的方法,其特征在于,将以所述预分裂时间点为时间片键范围起点的子数据块移动到另一分片中,包括:

基于移动操作指令的触发将以所述预分裂时间点为时间片键范围起点的子数据块移动到另一分片中。

7. 一种数据库更改写入分区的装置,其特征在于,所述装置包括:

获取单元,用于获取数据库当前分片中最新数据块的时间片键范围,所述最新数据块为当前写入数据的数据块;

确定单元,用于在所述获取单元获取的所述最新数据块的时间片键范围内确定预分裂时间点;

分裂单元,用于将所述最新数据块在所述确定单元确定的所述预分裂时间点进行分裂,生成两个子数据块;

移动单元,用于将所述分裂单元生成的以所述预分裂时间点为时间片键范围起点的子数据块移动到另一分片中。

8. 根据权利要求 7 所述的装置,其特征在于,所述获取单元包括:

监控模块,用于监控所述数据库分片集群中当前分片的数据量;

获取模块,用于当所述当前分片的数据量大于预设阈值时,获取所述当前分片中最新数据块的时间片键范围。

一种数据库更改写入分区的方法及装置

技术领域

[0001] 本发明涉及数据处理领域,特别是涉及一种数据库更改写入分区的方法及装置。

背景技术

[0002] 在 MongoDB 数据库的分片集群中,片键决定了每条数据应该被写入到集群中的哪个分片上,各个分片之间通过数据块的移动达到分片负载的均衡,因此在设置分片时,需要从数据集中选取一个键,用该键作为数据拆分的依据,这个键就是片键。片键的选择非常重要,MongoDB 会根据选定的片键将数据划分到有着相同片键的数据块中,而后这些数据块将根据片键的大致顺序分散到分片中。然而,当数据写入的负载较大且选择了单调递增片键如时间单调递增片键时,新的数据只会写入到时间最新的一个数据块中得到信息量巨大的数据块,导致数据的不平均分布。由于数据写入的负载较高,MongoDB 在各个分片中移动数据块的速度远远无法跟上数据的写入速度,而且一旦数据块的信息量大小超过一定数值,那么就无法在分片之间对其进行移动,从而使得大量数据堆积在一个分片上,导致硬盘的容量迅速达到极限。

发明内容

[0003] 有鉴于此,本发明提出了一种数据库更改写入分区的方法及装置,主要目的在于解决在数据库分片集群中选取单调递增片键进行分区时导致数据分片不均衡的问题。

[0004] 依据本发明的第一个方面,本发明提供了一种数据库更改写入分区的方法,包括:

[0005] 获取数据库当前分片中最新数据块的时间片键范围,最新数据块为当前写入数据的数据块;

[0006] 在最新数据块的时间片键范围内确定预分裂时间点;

[0007] 将最新数据块在预分裂时间点进行分裂,生成两个子数据块;

[0008] 将以预分裂时间点为时间片键范围起点的子数据块移动到另一分片中。

[0009] 依据本发明的第二个方面,本发明提供了一种数据库更改写入分区的装置,包括:

[0010] 获取单元,用于获取数据库当前分片中最新数据块的时间片键范围,最新数据块为当前写入数据的数据块;

[0011] 确定单元,用于在获取单元获取的最新数据块的时间片键范围内确定预分裂时间点;

[0012] 分裂单元,用于将最新数据块在选择单元选择的预分裂时间点进行分裂,生成两个子数据块;

[0013] 移动单元,用于将分裂单元生成的以预分裂时间点为时间片键范围起点的子数据块移动到另一分片中。

[0014] 借由上述技术方案,本发明实施例提供的数据库更改写入分区的方法及装置,能

够获取数据库当前分片中最新数据块的时间片键范围,并在该时间片键范围内确定预分裂时间点,将所述最新数据块在预分裂时间点进行分裂,生成两个子数据块,其中以预分裂时间点为时间片键范围起点的子数据块由于还未插入数据,因此将其移动到另一分片中,当实际时间到达预分裂时间点后,新的数据会自动写入到被移动的子数据块所在的新的分片中,从而避免了由于分裂前的最新数据块写入数据的速度高于数据块自动分裂和移动的速度导致当前分片的容量过高,确保了数据在数据库分片集群中的均衡分布。

[0015] 上述说明仅是本发明技术方案的概述,为了能够更清楚了解本发明的技术手段,而可依照说明书的内容予以实施,并且为了让本发明的上述和其它目的、特征和优点能够更明显易懂,以下特举本发明的具体实施方式。

附图说明

[0016] 通过阅读下文优选实施方式的详细描述,各种其他的优点和益处对于本领域普通技术人员将变得清楚明了。附图仅用于示出优选实施方式的目的,而并不认为是对本发明的限制。而且在整个附图中,用相同的参考符号表示相同的部件。在附图中:

[0017] 图 1 示出了本发明实施例提供了一种数据库更改写入分区的方法的流程图;

[0018] 图 2 示出了本发明实施例提供了一种数据库更改写入分区的装置的结构示意图;

[0019] 图 3 示出了本发明实施例提供的另一种数据库更改写入分区的装置的结构示意图。

具体实施方式

[0020] 下面将参照附图更加详细地描述本公开的示例性实施例。虽然附图中显示了本公开的示例性实施例,然而应当理解,可以以各种形式实现本公开而不应被这里阐述的实施例所限制。相反,提供这些实施例是为了能够更透彻地理解本公开,并且能够将本公开的范围完整的传达给本领域的技术人员。

[0021] 在使用数据库的业务中,随着业务的发展,产生的数据量也越来越大,因此基于数据分片的数据库构架应运而生。数据分片就是将整体数据分摊在多个存储设备上,这样每个存储设备的数据量相对变小,从而满足数据库性能需求。在设置分片时,需要从数据集合中选取一个键,用该键作为数据拆分的依据,这个键被称为片键。最初在数据库中插入数据时,数据会被插入到一个分片中的数据块中,随着数据块的数据量逐渐增大,该分片的负载逐渐增高,在这种情况下,数据库通常会自动将该数据块已写入数据的部分分裂出去并移动到其他分片中,从而保证各个分片的数据分布均衡。

[0022] 在本发明实施例中,由于采用时间单调递增片键作为数据拆分的依据,这样会导致数据一直被写入到最新的分片中,同时如果数据的写入负载过高,在各个分片间移动数据块的速度远远无法跟上数据的写入速度时,会导致当前分片的数据量增大,而且一旦数据块的信息量大小超过一定数值,那么就无法在分片之间对其进行移动,从而使得大量数据堆积在一个分片上,导致硬盘的容量迅速达到极限。

[0023] 为了解决在数据库分片集群中选取单调递增片键进行分区时导致数据分片不均衡的问题,本发明实施例提供了一种数据库更改写入分区的方法,如图 1 所示,该方法包括:

[0024] 101、获取数据库当前分片中最新数据块的时间片键范围。

[0025] 由于数据是被写入数据块中的,每个数据块所包含的数据范围是按照片键的取值来划分的,若对数据库进行分区,就需要对数据块进行操作。因此在本发明实施例中,需要执行步骤 101 获取数据库当前分片中最新数据块的时间片键范围,当前分片为数据库分片集群中当前正在写入数据的分片,其中,当前分片中最新数据块的时间片键范围为该分片中当前正在写入数据的数据块的时间片键范围。

[0026] 102、在最新数据块的时间片键范围内确定预分裂时间点。

[0027] 当获取到最新数据块的时间片键范围后,为了防止最新数据块自动分裂和移动的速度慢于数据写入速度导致的当前分片负载过高情况的发生,因此本发明实施例需要提前对最新数据块中未写入数据的部分进行分裂,得到未写入数据的空数据块,进行分裂的前提是需要知道分裂的节点,因此本发明实施例需要执行步骤 102 在最新数据块的时间片键范围内确定预分裂时间点。这里需要说明的是,选择的该预分裂时间点一定要确保在该预分裂时间点之后的数据块未被写入数据,也就是该预分裂时间点之后的数据块为空数据块。

[0028] 103、将最新数据块在预分裂时间点进行分裂,生成两个子数据块。

[0029] 当在步骤 102 中得到最新数据块的预分裂时间点后,需要在该预分裂时间点将最新数据块进行分裂操作,生成两个子数据块,其中一个子数据块为空数据块,未被写入数据,另一个子数据块为分裂前最新数据块中已写入数据的部分,将继续写入数据。

[0030] 104、将以预分裂时间点为时间片键范围起点的子数据块移动到另一分片中。

[0031] 当最新数据块在预分裂时间点被分裂后,需要将未被写入数据的子数据块,也就是以预分裂时间点为时间片键范围起点的子数据块移动到另一分片中,当系统时间到达预分裂时间点后,新的数据会被自动写入到以预分裂时间点为时间片键范围起点的子数据块所在的分片中,从而保证分裂前的最新数据块所在的分片的负载不会超过极限。

[0032] 本发明实施例提供的数据库更改写入分区的方法,能够获取数据库当前分片中最新数据块的时间片键范围,并在该时间片键范围内确定预分裂时间点,将所述最新数据块在预分裂时间点进行分裂,生成两个子数据块,其中以预分裂时间点为时间片键范围起点的子数据块由于还未插入数据,因此将其移动到另一分片中,当实际时间到达预分裂时间点后,新的数据会自动写入到被移动的子数据块所在的新的分片中,从而避免了由于分裂前的最新数据块写入数据的速度高于数据块自动分裂和移动的速度导致当前分片的容量过高,确保了数据在数据库分片集群中的均衡分布。

[0033] 进一步的,为了更好的对上述图 1 所示的方法进行理解,作为对上述实施方式的细化和扩展,本发明实施例将针对图 1 中的步骤进行详细说明。

[0034] 在本发明实施例中以时间单调递增片键为例,数据库分片集群中当前分片的最新数据块也就是当前被写入数据的数据块,例如当前被写入数据的数据块为最新数据块 M,其时间片键范围为 $[T_m, T_\infty)$, 则获取数据库当前分片中最新数据块的时间片键范围为 $[T_m, T_\infty)$ 。

[0035] 由于时间片键位于 $[T_m, T_\infty)$ 内的数据都被写入该最新数据块 M 中,因此随着时间的推移,该最新数据块 M 的容量为越来越大,导致其所在分片的容量逐渐接近极限。为了避免上述问题的发生,现有技术中数据库提供了自动拆分数据块的功能,将数据块中已经

写入数据的部分拆分出来并移动到其他分片中,从而防止当前分片的容量达到极限。但是数据库提供的自动拆分数据块的功能并不能拆分并移动数据块中未写入数据的部分,因此当数据写入该最新数据块 M 的速度远大于数据库自动拆分该最新数据块 M 并将拆分后的数据块移动到其他分片中的速度时,该分片的容量将迅速达到极限。为了解决上述问题,本发明实施例需要将该最新数据块 M 中未写入数据的部分提前进行拆分,获取该最新数据块 M 中未写入数据的部分并将其进行移动,从而保证当前分片的容量不会达到极限。

[0036] 作为一种可选的实施方式,本发明实施例会实时监控数据库分片集群中当前分片的数据量,当当前分片的数据量大于预设阈值时,会获取当前分片中最新数据块 M 的时间片键范围 $[T_m, T_\infty)$,并在该时间片键范围内确定预分裂时间点,从而将最新数据块 M 在预分裂时间点进行拆分。

[0037] 由于本发明实施例的关键在于拆分并移动最新数据块中未写入数据的部分,因此在最新数据块 M 的时间片键范围 $[T_m, T_\infty)$ 内确定的预分裂时间点必须能够将该最新数据块 M 分裂为两个部分,其中一部分从未写入数据。作为一种可选的实施方式,本发明实施例可以将当前时间点 T_{now} 之后的一个时间点 T_{future} 确定为预分裂时间点。由于在当前时间点 T_{now} 处,数据写入状态只有两种,一种为正在写入数据,一种为未写入数据,如果在当前时间点 T_{now} 处对最新数据块 M 进行分裂,则很有可能被正在写入数据的情况所干扰,导致无法正确对最新数据块 M 进行分裂。而在当前时间点 T_{now} 之后的时间点 T_{future} 处的数据写入状态只有一种情况,即未写入数据状态。因此在时间点 T_{future} 处一定能够将最新数据块 M 进行分裂得到未写入数据的部分。

[0038] 当在最新数据块 M 的时间片键范围 $[T_m, T_\infty)$ 内确定预分裂时间点 T_{future} 后,就可以在预分裂时间点 T_{future} 处将最新数据块 M 进行分裂,由于数据库无法对空数据块也就是数据块中未写入数据的部分进行自动分裂,因此,本发明实施例提供了一种可选的实施方式,由数据库基于分裂操作指令的触发将最新数据块 M 在预分裂时间点 T_{future} 处进行分裂。在具体实施过程中,可以使用 `sp split` 命令在预分裂时间点 T_{future} 处将最新数据块 M 进行分裂,生成两个子数据块,子数据块 A 和子数据块 B。其中,两个子数据块具有各自新的时间片键范围,若子数据块 B 为未写入数据的部分,那么子数据块 B 的时间片键范围为 $[T_{future}, T_\infty)$,以预分裂时间点 T_{future} 和分裂前的最新数据块 M 的时间片键范围的终点 T_∞ 作为自身时间片键范围的起点和终点;子数据块 A 的时间片键范围为 $[T_m, T_{future})$,以分裂前的最新数据块 M 的时间片键范围的起点 T_m 和预分裂时间点 T_{future} 作为自身时间片键范围的起点和终点。

[0039] 当将最新数据块 M 在预分裂时间点 T_{future} 处进行分裂得到未写入数据的子数据块 B 后,需要将子数据块 B 移动到另一分片中,从而避免分裂前最新数据块 M 所在分片的容量达到极限。由于数据库无法对分裂后得到的空数据块也就是数据块中未写入数据的部分进行自动迁移,因此,本发明实施例提供了一种可选的实施方式,由数据库基于移动操作指令的触发将以预分裂时间点 T_{future} 为时间片键范围起点的子数据块 B 移动到另一分片中。在具体实施过程中,可以使用 `moveChunk` 命令来迁移子数据块 B,将其移动到另一分片中。当将子数据块 B 移动到另一分片中后,若实际时间在到达预分裂时间点 T_{future} 之前,新的数据会被数据库继续自动写入子数据块 A 所在的分片中;若实际时间到达预分裂时间点 T_{future} 之后,新的数据会被数据库自动写入子数据块 B 所在的分片中,从而避免分裂前最新数据

块 M 所在分片的容量达到极限。

[0040] 本发明实施例通过监控数据库分片集群中当前分片的数据量,当当前分片的数据量大于预设阈值时,触发对最新数据块的预分裂和迁移,从而避免最新数据块所在分片的容量达到极限。此外,通过多次在未来时间点触发数据块的分裂和迁移操作,能够均衡的在数据库分片集群中分布写入的数据。

[0041] 作为对上述图 1 所示方法的应用,本发明实施例提供了一种数据库更改写入分区的装置,如图 2 所示,该装置包括:获取单元 21、确定单元 22、分裂单元 23 及移动单元 24,其中,

[0042] 获取单元 21,用于获取数据库当前分片中最新数据块的时间片键范围,最新数据块为当前写入数据的数据块;

[0043] 确定单元 22,用于在获取单元 21 获取的最新数据块的时间片键范围内确定预分裂时间点;

[0044] 分裂单元 23,用于将最新数据块在确定单元 22 确定的预分裂时间点进行分裂,生成两个子数据块;

[0045] 移动单元 24,用于将分裂单元 23 生成的以预分裂时间点为时间片键范围起点的子数据块移动到另一分片中。

[0046] 进一步的,如图 3 所示,获取单元 21 包括:

[0047] 监控模块 211,用于监控数据库分片集群中当前分片的数据量;

[0048] 获取模块 212,用于当当前分片的数据量大于预设阈值时,获取当前分片中最新数据块的时间片键范围。

[0049] 进一步的,确定单元 22 用于在最新数据块的时间片键范围内,将当前时间点之后的任一个时间点确定为预分裂时间点。

[0050] 进一步的,分裂单元 23 用于基于分裂操作指令的触发将最新数据块在预分裂时间点进行分裂。

[0051] 进一步的,移动单元 24 用于基于移动操作指令的触发将以预分裂时间点为时间片键范围起点的子数据块移动到另一分片中。

[0052] 本发明实施例提供的数据库更改写入分区的装置,能够获取数据库当前分片中最新数据块的时间片键范围,并在该时间片键范围内确定预分裂时间点,将所述最新数据块在预分裂时间点进行分裂,生成两个子数据块,其中以预分裂时间点为时间片键范围起点的子数据块由于还未插入数据,因此将其移动到另一分片中,当实际时间到达预分裂时间点后,新的数据会自动写入到被移动的子数据块所在的新的分片中,从而避免了由于分裂前的最新数据块写入数据的速度高于数据块自动分裂和移动的速度导致当前分片的容量过高,确保了数据在数据库分片集群中的均衡分布。

[0053] 此外,本发明实施例通过实时监控数据库分片集群中当前分片的数据量,当当前分片的数据量大于预设阈值时,触发对最新数据块的预分裂和迁移,从而避免最新数据块所在分片的容量达到极限。同时,通过多次在未来时间点触发数据块的分裂和迁移操作,能够均衡的在数据库分片集群中分布写入的数据。

[0054] 在上述实施例中,对各个实施例的描述都各有侧重,某个实施例中未详述的部分,可以参见其他实施例的相关描述。

[0055] 可以理解的是,上述方法及装置中的相关特征可以相互参考。另外,上述实施例中的“第一”、“第二”等是用于区分各实施例,而并不代表各实施例的优劣。

[0056] 所属领域的技术人员可以清楚地了解到,为描述的方便和简洁,上述描述的系统、装置和单元的具体工作过程,可以参考前述方法实施例中的对应过程,在此不再赘述。

[0057] 在此提供的算法和显示不与任何特定计算机、虚拟系统或者其它设备固有相关。各种通用系统也可以与基于在此的示教一起使用。根据上面的描述,构造这类系统所要求的结构是显而易见的。此外,本发明也不针对任何特定编程语言。应当明白,可以利用各种编程语言实现在此描述的本发明的内容,并且上面对特定语言所做的描述是为了披露本发明的最佳实施方式。

[0058] 在此处所提供的说明书中,说明了大量具体细节。然而,能够理解,本发明的实施例可以在没有这些具体细节的情况下实践。在一些实例中,并未详细示出公知的方法、结构和技术,以便不模糊对本说明书的理解。

[0059] 类似地,应当理解,为了精简本公开并帮助理解各个发明方面中的一个或多个,在上面对本发明的示例性实施例的描述中,本发明的各个特征有时被一起分组到单个实施例、图、或者对其的描述中。然而,并不应将该公开的方法解释成反映如下意图:即所要求保护的本发明要求比在每个权利要求中所明确记载的特征更多的特征。更确切地说,如下面的权利要求书所反映的那样,发明方面在于少于前面公开的单个实施例的所有特征。因此,遵循具体实施方式的权利要求书由此明确地并入该具体实施方式,其中每个权利要求本身都作为本发明的单独实施例。

[0060] 本领域那些技术人员可以理解,可以对实施例中的设备中的模块进行自适应性地改变并且把它们设置在与该实施例不同的一个或多个设备中。可以把实施例中的模块或单元或组件组合成一个模块或单元或组件,以及此外可以把它分成多个子模块或子单元或子组件。除了这样的特征和/或过程或者单元中的至少一些是相互排斥之外,可以采用任何组合对本说明书(包括伴随的权利要求、摘要和附图)中公开的所有特征以及如此公开的任何方法或者设备的所有过程或单元进行组合。除非另外明确陈述,本说明书(包括伴随的权利要求、摘要和附图)中公开的每个特征可以由提供相同、等同或相似目的的替代特征来代替。

[0061] 此外,本领域的技术人员能够理解,尽管在此所述的一些实施例包括其它实施例中所包括的某些特征而不是其它特征,但是不同实施例的特征的组合意味着处于本发明的范围之内并且形成不同的实施例。例如,在下面的权利要求书中,所要求保护的实施例的任意之一都可以以任意的组合方式来使用。

[0062] 本发明的各个部件实施例可以以硬件实现,或者以在一个或者多个处理器上运行的软件模块实现,或者以它们的组合实现。本领域的技术人员应当理解,可以在实践中使用微处理器或者数字信号处理器(DSP)来实现根据本发明实施例的发明名称(如确定网站内链接等级的装置)中的一些或者全部部件的一些或者全部功能。本发明还可以实现为用于执行这里所描述的方法的一部分或者全部的设备或者装置程序(例如,计算机程序和计算机程序产品)。这样的实现本发明的程序可以存储在计算机可读介质上,或者可以具有一个或者多个信号的形式。这样的信号可以从因特网网站上下载得到,或者在载体信号上提供,或者以任何其他形式提供。

[0063] 应该注意的是上述实施例对本发明进行说明而不是对本发明进行限制,并且本领域技术人员在不脱离所附权利要求的范围的情况下可设计出替换实施例。在权利要求中,不应将位于括号之间的任何参考符号构造成对权利要求的限制。单词“包含”不排除存在未列在权利要求中的元件或步骤。位于元件之前的单词“一”或“一个”不排除存在多个这样的元件。本发明可以借助于包括有若干不同元件的硬件以及借助于适当编程的计算机来实现。在列举了若干装置的单元权利要求中,这些装置中的若干个可以是通过同一个硬件项来具体体现。单词第一、第二、以及第三等的使用不表示任何顺序。可将这些单词解释为名称。

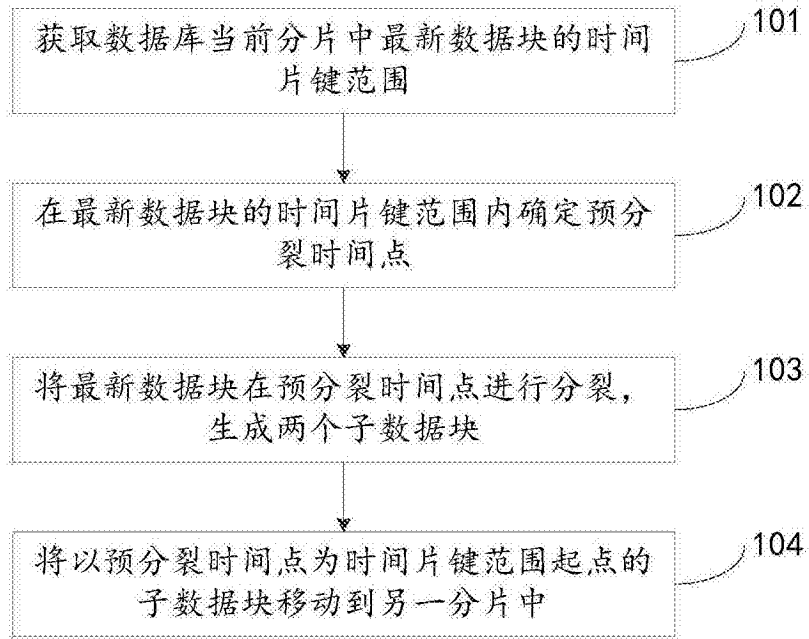


图 1

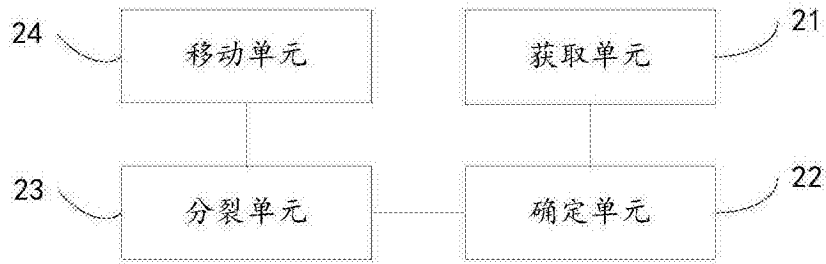


图 2

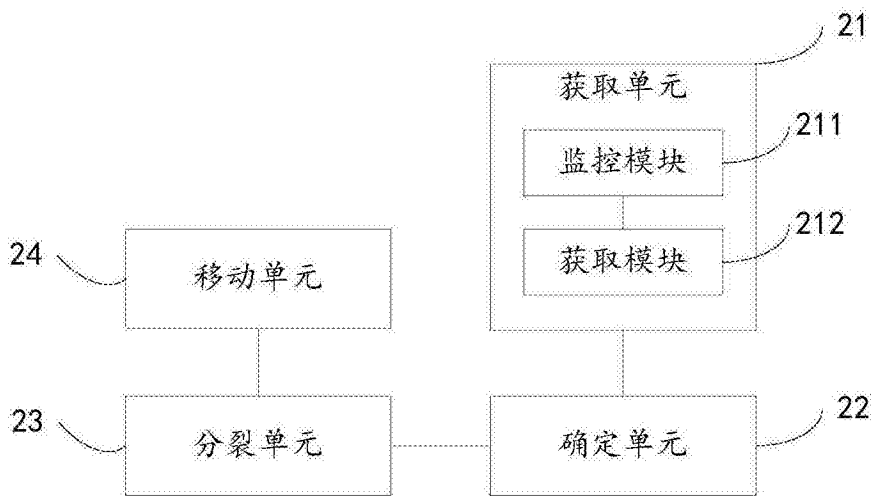


图 3