(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 108563697 B
(45)授权公告日 2021.02.26

(21)申请号 201810239892.9
(22)申请日 2018.03.22
(65)同一申请的已公布的文献号
申请公布号 CN 108563697 A
(43)申请公布日 2018.09.21
(73)专利权人 腾讯科技(深圳)有限公司
地址 518044 广东省深圳市南山区高新区科技中一路腾讯大厦35层
专利权人 腾讯云计算(北京)有限责任公司
(72)发明人 严俊明
(74)专利代理机构 北京同达信恒知识产权代理有限公司 11291
代理人 郭润湘
(51)Int.Cl.
G06F 16/22(2019.01)

审查员 尤晓美

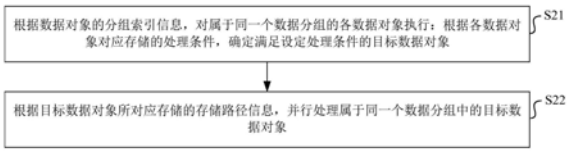
权利要求书2页 说明书12页 附图3页

(54)发明名称

一种数据处理方法、装置和存储介质

(57)摘要

本发明公开了一种数据处理方法、装置和存储介质，用以快速定位待处理的文件，提高数据处理的时效性。所述数据处理方法，包括：根据数据对象的分组索引信息，对属于同一个数据分组的各数据对象执行：根据各数据对象对应存储的处理条件，确定满足设定处理条件的目标数据对象；并根据目标数据对象所对应存储的存储路径信息，并行处理属于同一个数据分组中的所有目标数据对象。



CN 108563697 B

1.一种数据处理方法,其特征在于,包括:

根据数据对象的分组索引信息,对属于同一个数据分组的各数据对象执行:根据各数据对象对应存储的处理条件,确定满足设定处理条件的目标数据对象;并

根据所述目标数据对象所对应存储的存储路径信息,并行处理属于同一个数据分组中的所述目标数据对象;

其中,所述分组索引信息由主索引信息和从索引信息组成;以及

按照以下方法确定所述数据对象的分组索引信息:

确定用户标识中任两位相邻位置的数值为所述主索引信息;

针对根据所述主索引信息确定出的数据分组所包含的数据对象,进一步根据数据对象的存储路径信息确定数据对象的从索引信息;

所述主索引信息用于根据用户标识对待处理的数据对象进行一次散列处理,所述从索引信息用于根据存储路径信息对待处理的数据对象进行二次散列处理。

2.如权利要求1所述的方法,其特征在于,所述分组索引信息根据所述数据对象对应的用户标识确定;或者所述分组索引信息根据所述数据对象的存储路径信息确定。

3.如权利要求1所述的方法,其特征在于,所述从索引信息为根据所述数据对象的存储路径确定出的MD5值确定。

4.如权利要求3所述的方法,其特征在于,所述分组索引信息中还包括数据处理规则和存储空间标识,以及所述数据对象为满足所述数据处理规则的数据对象,并且每一个存储空间标识对应一种数据处理规则。

5.如权利要求4所述的方法,其特征在于,采用三级索引结构存储所述分组索引信息,其中,每一级索引由散列索引和分区索引组成,一级索引的散列索引中存储有业务标识,一级索引的分区索引中存储有所述数据对象对应的用户标识和存储空间标识,一级索引的值针对不同的存储空间标识存储其对应的数据处理规则;二级索引的散列索引中存储有根据所述用户标识中任两位相邻位置的数值生成所述分组索引信息,二级索引的分区索引中存储有根据所述数据对象的存储路径确定出的MD5值确定所述数据对象对应的分组索引信息;三级索引的散列索引中存储有根据所述数据对象的存储路径确定出的MD5值确定所述数据对象对应的分组索引信息;三级索引的分区索引中存储有所述数据对象对应的处理条件和存储路径信息。

6.如权利要求5所述的方法,其特征在于,所述处理条件包括处理时间信息;以及

所述方法,还包括:

根据所述处理时间信息,按照处理时间的先后顺序对三级索引的分区索引进行排序。

7.一种数据处理装置,其特征在于,包括:

确定单元,用于根据数据对象的分组索引信息,对属于同一个数据分组的各数据对象执行:根据各数据对象对应存储的处理条件,确定满足设定处理条件的目标数据对象;

处理单元,用于根据所述目标数据对象所对应存储的存储路径信息,并行处理属于同一个数据分组中的所述目标数据对象;

其中,所述分组索引信息由主索引信息和从索引信息组成;以及

所述确定单元,还用于确定用户标识中任两位相邻位置的数值为所述主索引信息;以及针对根据所述主索引信息确定出的数据分组所包含的数据对象,进一步根据数据对象的

存储路径信息确定数据对象的从索引信息；

所述主索引信息用于根据用户标识对待处理的数据对象进行一次散列处理，所述从索引信息用于根据存储路径信息对待处理的数据对象进行二次散列处理。

8.如权利要求7所述的装置，其特征在于，所述分组索引信息根据所述数据对象对应的用户标识确定；或者所述分组索引信息根据所述数据对象的存储路径信息确定。

9.如权利要求7所述的装置，其特征在于，所述主索引信息为所述用户标识中任两位相邻位置的数值；以及所述从索引信息为根据所述数据对象的存储路径确定出的MD5值确定。

10.如权利要求9所述的装置，其特征在于，其特征在于，所述分组索引信息中还包括数据处理规则和存储空间标识，以及所述数据对象为满足所述数据处理规则的数据对象，并且每一个存储空间标识对应一种数据处理规则。

11.如权利要求10所述的装置，其特征在于，采用三级索引结构存储所述分组索引信息，其中，每一级索引由散列索引和分区索引组成，一级索引的散列索引中存储有业务标识，一级索引的分区索引中存储有所述数据对象对应的用户标识和存储空间标识，一级索引的值针对不同的存储空间标识存储其对应的数据处理规则；二级索引的散列索引中存储有根据所述用户标识中任两位相邻位置的数值生成所述分组索引信息，二级索引的分区索引中存储有根据所述数据对象的存储路径确定出的MD5值确定所述数据对象对应的分组索引信息；三级索引的散列索引中存储有根据所述数据对象的存储路径确定出的MD5值确定所述数据对象对应的分组索引信息；三级索引的分区索引中存储有所述数据对象对应的处理条件和存储路径信息。

12.一种计算装置，其特征在于，包括至少一个处理器、以及至少一个存储器，其中，所述存储器存储有计算机程序，当所述程序被所述处理器执行时，使得所述处理器执行权利要求1~6任一权利要求所述方法的步骤。

13.一种计算机可读介质，其特征在于，其存储有可由终端设备执行的计算机程序，当所述程序在终端设备上运行时，使得所述终端设备执行权利要求1~6任一所述方法的步骤。

一种数据处理方法、装置和存储介质

技术领域

[0001] 本发明涉及数据处理技术领域，尤其涉及一种数据处理方法、装置和存储介质。

背景技术

[0002] 本部分旨在为权利要求书中陈述的本发明的实施方式提供背景或上下文。此处的描述不因为包括在本部分中就承认是现有技术。

[0003] 云对象存储(Cloud Object Storage，COS)是一种能够存储海量数据的分布式存储服务，用户可以上传任意数量的文件、视频、图片等对象，同时提供高效的下载访问服务来获取对象，实现了通过互联网随时对大量数据进行批量存储和处理。

[0004] COS系统中存储的文件数量达到了万亿量级，当存储的文件超过一段时间不再被访问时，通常需要从COS系统中删除这些文件，例如日志数据文件和监控数据文件等等。现有的过期文件删除方法主要采用全量扫描法。这种方法按照一定的周期定时扫描COS系统中的全量文件，针对每一文件，判断其是否满足过期删除规则，并记录满足规则的文件，待扫描工作结束后，统一删除记录的文件。

[0005] 由于COS系统中存储有海量的文件，扫描全量文件列表耗时较长，对系统资源消耗较大，无法满足快速定位和删除大规模文件的需求，而且随着COS系统中存储的文件数量的增加，该方法消耗的系统资源越来越多，处理时效越来越差。

发明内容

[0006] 本发明实施例提供一种数据处理方法、装置和存储介质，用以快速定位待处理的文件，提高数据处理的时效性。

[0007] 提供一种数据处理方法，包括：

[0008] 根据数据对象的分组索引信息，对属于同一个数据分组的各数据对象执行：根据各数据对象对应存储的处理条件，确定满足设定处理条件的目标数据对象；并

[0009] 根据所述目标数据对象所对应存储的存储路径信息，并行处理属于同一个数据分组中的所述目标数据对象。

[0010] 可选地，所述分组索引信息根据所述数据对象对应的用户标识确定；或者所述分组索引信息根据所述数据对象的存储路径信息确定。

[0011] 可选地，所述分组索引信息由主索引信息和从索引信息组成；以及

[0012] 按照以下方法确定数据对象的分组索引信息：

[0013] 根据所述数据对象对应的用户标识确定主索引信息；

[0014] 针对根据所述主索引信息确定出的数据分组所包含的数据对象，进一步根据数据对象的存储路径信息确定数据对象的从索引信息。

[0015] 可选地，所述主索引信息为所述用户标识中任两位相邻位置的数值；以及所述从索引信息为根据所述数据对象的存储路径确定出的MD5值确定。

[0016] 可选地，所述分组索引信息中还包括数据处理规则和存储空间标识，以及所述数

据对象为满足所述数据处理规则的数据对象，并且每一个存储空间标识对应一种数据处理规则。

[0017] 可选地，采用三级索引结构存储所述分组索引信息，其中，每一级索引由散列索引和分区索引组成，一级索引的散列索引中存储有业务标识，一级索引的分区索引中存储有所述数据对象对应的用户标识和存储空间标识，一级索引的值针对不同的存储空间标识存储其对应的数据处理规则；二级索引的散列索引中存储有根据所述用户标识中任两位相邻位置的数值生成所述分组索引信息，二级索引的分区索引中存储有根据所述数据对象的存储路径确定出的MD5值确定所述数据对象对应的分组索引信息；三级索引的散列索引中存储有根据所述数据对象的存储路径确定出的MD5值确定所述数据对象对应的分组索引信息；三级索引的分区索引中存储有所述数据对象对应的处理条件和存储路径信息。

[0018] 可选地，所述处理条件包括处理时间信息；以及

[0019] 所述方法，还包括：

[0020] 根据所述处理时间信息，按照处理时间的先后顺序对三级索引的分区索引进行排序。

[0021] 还提供一种数据处理装置，包括：

[0022] 确定单元，用于根据数据对象的分组索引信息，对属于同一个数据分组的各数据对象执行：根据各数据对象对应存储的处理条件，确定满足设定处理条件的目标数据对象；

[0023] 处理单元，用于根据所述目标数据对象所对应存储的存储路径信息，并行处理属于同一个数据分组中的所述目标数据对象。

[0024] 可选地，所述分组索引信息根据所述数据对象对应的用户标识确定；或者所述分组索引信息根据所述数据对象的存储路径信息确定。

[0025] 可选地，所述分组索引信息由主索引信息和从索引信息组成；以及

[0026] 所述确定单元，还用于根据所述数据对象对应的用户标识确定数据对象的主索引信息；

[0027] 针对根据所述主索引信息确定出的数据分组所包含的数据对象，进一步根据数据对象的存储路径信息确定数据对象的从索引信息。

[0028] 可选地，所述主索引信息为所述用户标识中任两位相邻位置的数值；以及所述从索引信息为根据所述数据对象的存储路径确定出的MD5值确定。

[0029] 可选地，所述分组索引信息中还包括数据处理规则和存储空间标识，以及所述数据对象为满足所述数据处理规则的数据对象，并且每一个存储空间标识对应一种数据处理规则。

[0030] 可选地，采用三级索引结构存储所述分组索引信息，其中，每一级索引由散列索引和分区索引组成，一级索引的散列索引中存储有业务标识，一级索引的分区索引中存储有所述数据对象对应的用户标识和存储空间标识，一级索引的值针对不同的存储空间标识存储其对应的数据处理规则；二级索引的散列索引中存储有根据所述用户标识中任两位相邻位置的数值生成所述分组索引信息，二级索引的分区索引中存储有根据所述数据对象的存储路径确定出的MD5值确定所述数据对象对应的分组索引信息；三级索引的散列索引中存储有根据所述数据对象的存储路径确定出的MD5值确定所述数据对象对应的分组索引信息；三级索引的分区索引中存储有所述数据对象对应的处理条件和存储路径信息。

[0031] 可选地，所述处理条件包括处理时间信息；以及

[0032] 所述装置，还包括：

[0033] 排序单元，用于根据所述处理时间信息，按照处理时间的先后顺序对三级索引的分区索引进行排序。

[0034] 还提供一种计算装置，包括至少一个处理器、以及至少一个存储器，其中，所述存储器存储有计算机程序，当所述程序被所述处理器执行时，使得所述处理器执行上述数据处理方法所述的步骤。

[0035] 还提供一种计算机可读介质，其存储有可由终端设备执行的计算机程序，当所述程序在终端设备上运行时，使得所述终端设备执行上述数据处理方法所述的步骤。

[0036] 本发明实施例提供的数据处理方法、装置和存储介质，通过将带处理的数据对象划分为不同的数据分组，在每一数据处理周期，分别轮询每一分组，确定出满足处理条件的目标数据对象，并根据针对目标对象所对应存储的存储路径信息快速定位到待处理的目标数据对象，并发处理同一数据分组中包含的所有目标数据对象，从而提高了数据处理的时效性。

[0037] 本发明的其它特征和优点将在随后的说明书中阐述，并且，部分地从说明书中变得显而易见，或者通过实施本发明而了解。本发明的目的和其他优点可通过在所写的说明书、权利要求书、以及附图中所特别指出的结构来实现和获得。

附图说明

[0038] 此处所说明的附图用来提供对本发明的进一步理解，构成本发明的一部分，本发明的示意性实施例及其说明用于解释本发明，并不构成对本发明的不当限定。在附图中：

[0039] 图1为根据本发明实施方式的应用场景示意图；

[0040] 图2为根据本发明实施方式的数据处理方法的实施流程示意图；

[0041] 图3为根据本发明另一实施方式的数据处理方法的实施流程示意图；

[0042] 图4为根据本发明实施例方式的数据处理装置的结构示意图；

[0043] 图5为根据本发明实施方式的计算装置的结构示意图。

具体实施方式

[0044] 为了从海量的数据对象中定位待处理的数据对象，提高数据对象的处理时效，本发明实施例提供了一种数据处理方法、装置和存储介质。

[0045] 首先，对本发明实施例中涉及的部分用语进行说明，以便于本领域技术人员理解。

[0046] 存储桶即Bucket，在COS系统中用于存储对象。一个存储桶中可以存储多个对象。存储桶名由用户自定义的字符串和系统自动生成的数字串用中划线链接而成，以保证该存储桶全球唯一。

[0047] 对象即Object，COS中存储的基本单元。

[0048] APPID是用户云账户的账户标识之一，用于关联云资源。在用户成功申请云账户后，系统自动为用户分配一个APPID。

[0049] KV(key-value)存储系统，KV存储系统中的key由两部分组成：key_hash和key_range，key_hash用于对key值进行散列，提高并发性能。key_range用于实现有序排列。

[0050] 以下结合说明书附图对本发明的优选实施例进行说明，应当理解，此处所描述的优选实施例仅用于说明和解释本发明，并不用于限定本发明，并且在不冲突的情况下，本发明中的实施例及实施例中的特征可以相互组合。

[0051] 如图1所示，其为本法实施例提供的数据处理方法的应用场景示意图。用户10通过终端设备11中安装的客户端登录云服务器12，其中，客户端可以为网页的浏览器，也可以为安装于终端设备，如手机，平板电脑等中的客户端。

[0052] 终端设备11与云服务器12之间通过网络进行通信连接，该网络可以为局域网、广域网等。终端设备11可以为便携设备(例如:手机、平板、笔记本电脑等)，也可以为个人电脑(PC,Personal Computer)，云服务器12可以为任何能够提供互联网服务的设备。

[0053] 其中，用户10利用终端设备11通过向云服务器12注册获得用户名，云服务器12在用户进行注册成功后存储用户名以及与用户10设置的用户密码作为认证信息，后续用户10利用终端设备11再次登录云服务器12时，云服务器12向客户端返回登录页面，用户在客户端显示的登录页面输入认证信息(即用户名和用户密码)并提交给云服务器12，云服务器12比较用户提交认证信息与自身在用户注册时存储的认证信息是否一致以确定是否允许用户登录。用户注册成功后，系统自动为用户分配APPID。

[0054] 在用户10注册成功后，便可以创建存储桶用于存储数据对象。针对不同的业务类型，同一用户可以创建多个存储桶分别存储。针对不同的存储桶，用户可以设置不同的数据处理规则，例如，针对视频监控业务，用户可以创建存储桶1用于存储视频文件，设置该存储桶对应的数据处理规则为30天有效，即视频文件的存储时间超过30天即可删除。而针对文档存储类业务，用户可以创建存储桶2用于存储各类文档文件，设置该存储桶对应的数据处理规则为7天有效，即文档文件的存储时间超过7天即可删除。

[0055] 本发明实施例提供的数据处理方法可以应用于云服务器12中，云服务器12根据本发明实施例提供的数据处理方法对其存储的数据对象进行处理。需要说明的是，本发明实施例中涉及的处理操作可以包括任意对云服务器中存储的数据对象的处理操作，例如，删除操作，修改操作等等。

[0056] 下面结合图1的应用场景，参考图2来描述根据本发明示例性实施方式的数据处理方法。需要注意的是，上述应用场景仅是为了便于理解本发明的精神和原理而示出，本发明的实施方式在此方面不受任何限制。相反，本发明的实施方式可以应用于适用的任何场景。

[0057] 本发明实施例中，为了快速定位待处理的数据对象，提高数据处理的时效性，本发明实施例中，通过构造分组索引信息来快速定位待处理的处理对象，并根据分组索引信息针对属于同一数据分组的待处理的数据对象进行并行处理。

[0058] 在一个实施例中，分组索引信息可以根据用户标识(APPID)确定，例如，根据APPID中任N位相同位置的数值确定，其中，N为大于1的自然数，如根据APPID的最后三位确定，或者根据APPID的前三位确定，又如，根据APPID的最后两位确定等等，本发明实施例对此不进行限定。这样，根据APPID可以将用户划分为不同的数据分组并行进行处理。

[0059] 在另一实施例中，分组索引信息还可以根据数据对象的存储路径信息确定，例如，计算每一数据对象的存储路径的MD5(信息摘要)值，根据MD5中任M位相同位置的数值确定分组索引信息，其中，M为大于1的自然数。如根据MD5值最后三位确定分组索引信息确定分组索引信息，或者根据MD5值前三位确定分组索引信息，又如，根据MD5值最后两位确定分组

索引信息等等，本发明实施例对此不进行限定。这样，可以根据数据对象的存储路径将数据对象划分为不同的数据分组进行处理。

[0060] 具体实施时，为了避免同一用户待处理的数据对象数量较多，在同一时间内只能处理同一用户的数据对象，而影响其它用户符合处理条件的数据对象的处理，本发明实施例中，可以结合用户标识和数据对象的文件存储路径对数据对象进行分组处理。在这种实施方式中，在又一实施例中，分组索引信息还可以由两部分组成，分别为主索引信息和从索引信息，其中，主索引信息可以根据数据对象对应的APPID确定，例如，根据APPID的最后三位确定，或者根据APPID的前三位确定，又如，根据APPID的最后两位确定等等，进一步地，可以根据数据对象的存储路径信息确定从索引信息。同样，可以计算每一数据对象的存储路径的MD5(信息摘要)值，根据MD5中任M位相同位置的数值确定分组索引信息，如根据MD5值最后三位确定分组索引信息确定分组索引信息，或者根据MD5值前三位确定分组索引信息，又如，根据MD5值最后两位确定分组索引信息等等。这样，在具体实施时，可以首先根据数据对象对应的用户标识确定数据对象的主索引信息，这样可以将待处理的数据对象根据用户标识进行一次散列；在根据用户标识对数据对象进行散列之后，还可以针对根据所述主索引信息确定出的数据分组所包含的数据对象，进一步根据数据对象的存储路径信息确定数据对象的从索引信息，即针对根据用户标识确定出的属于同一数据分组的数据对象再次根据存储路径信息对数据对象进行二次散列处理，由此，可以将同一用户的不同数据对象分散在不同的数据分组中进行处理。

[0061] 需要说明的是，本发明实施例中只是将待处理的数据对象划分为不同的数据分组以并行处理同一数据分组内的满足处理条件的目标数据对象，而不是在数据存储时进行分组存储，数据存储仍然按照现有的方式进行存储。

[0062] 如图2所示，其为本发明实施例提供的数据处理方法的实施流程示意图，可以包括以下步骤：

[0063] S21、根据数据对象的分组索引信息，对属于同一个数据分组的各数据对象执行：根据各数据对象对应存储的处理条件，确定满足设定处理条件的目标数据对象。

[0064] 具体实施时，可以定期对数据对象进行处理，例如，每隔一定的处理周期对数据对象进行一次轮询。其中，数据对象的处理周期可以根据实际需要进行设置，例如，可以设置每天为一个处理周期，也可以设置每K个小时为一个处理周期，K为大于等于1的自然数。

[0065] 在每一处理周期开始时间到达时，云服务器根据分组索引信息分别轮询每一数据分组，针对每一数据分组，遍历该数据分组包含的每一数据对象，根据该数据对象对应的处理时间信息选择出处理时间到达的目标数据对象。

[0066] 具体实施时，本发明实施例中涉及的处理条件可以为处理时间信息。

[0067] S22、根据目标数据对象所对应存储的存储路径信息，并行处理属于同一个数据分组中的目标数据对象。

[0068] 针对每一数据分组，根据步骤S21选择出的目标数据对象并发进行处理。

[0069] 具体实施时，如果某一用户存储的数据对象数量较多，则可能导致在单位时间内处理的数据对象集中于某个APPID，引发数据处理瓶颈，为了解决该问题，本发明实施例中，可以针对APPID进行散列处理，用于离散待处理的数据对象对应的APPID，或者根据数据对象的存储路径确定分组索引信息。进一步地，如果根据数据对象对应的用户标识确定分组

索引信息，则还可以进一步根据数据对象的存储路径进一步将待处理的数据对象进行离散。这样，可以将不同用户的不同的数据对象离散在不同的数据分组中。

[0070] 具体实施时，分组索引信息中还可以包括数据处理规则和存储空间标识，例如，该存储空间标识可以为存储桶标识，以及所述数据对象为满足所述数据处理规则的数据对象，并且每一个存储空间标识对应一种数据处理规则。这样，在接收到数据对象存储请求时，即在用户上传一个新的数据对象时，首先根据APPID和存储桶标识，判断是否有作用于该数据对象的数据处理规则，如果有，则进一步根据数据对象对应的用户标识和/或存储路径标识生成该数据对象的分组索引信息添加到索引表中，以便后续快速定位到该数据对象。

[0071] 具体实施时，本发明实施例中采用KV存储系统存储各数据对象的分组索引信息。具体地，可以采用三级索引结构存储各数据对象的分组索引信息。每一级索引由散列索引(key_hash)和分区索引(key_range)组成，其中，一级索引的散列索引中存储有业务标识，一级索引的分区索引中存储有所述数据对象对应的用户标识和存储空间标识，一级索引的值针对不同的存储空间标识存储其对应的数据处理规则；二级索引的散列索引中存储有根据所述用户标识中任两位相邻位置的数值生成所述分组索引信息，二级索引的分区索引中存储有根据所述数据对象的存储路径确定出的MD5值确定所述数据对象对应的分组索引信息；三级索引的散列索引中存储有根据所述数据对象的存储路径确定出的MD5值确定所述数据对象对应的分组索引信息；三级索引的分区索引中存储有所述数据对象对应的处理条件和存储路径信息，如表1所示，其为三级索引结构一种可能的数据结构示意：

[0072] 表1

[0073]

索引类型	key_hash	key_range
一级索引	/op	/appid/bucket
二级索引	/op/appid_prefix	/appid/bucket/rule_ctime/hashkey/appid_prefix
三级索引	/appid/bucket/rule_ctime/hashkey/appid_prefix	/processtime/filepath

[0074] 其中，一级索引存储业务的规则集合，key_hash存储的是op字段代表各个业务，例如，其可以为COS数据对象存储业务，key_range存储的是用户标识和存储桶标识，一级索引的key对应的值存储针对每个bucket设置的过期规则。二级索引用于打散待处理的数据对象，可以提高数据对象处理性能。其中，appid_prefix，其可以取appid中的任两位相邻位置的数值，例如，可以取appid的最后两位，用于离散待处理的数据对象，避免单位时间内分发的数据对象处理任务集中在某个appid。rule_ctime代表数据处理规则的创建时间，用于识别当前的数据记录是否有效。

[0075] 具体实施时，为了避免在单位时间内集中于某一用户的数据对象，本发明实施例中还可以进一步对数据对象进行散列处理，本发明实施例中可以根据数据对象的存储路径计算其对应的MD5值，根据MD5值对数据对象进行散列处理，例如可以根据MD5值中任R位连续的数值进行散列，R为大于1的自然数，如根据MD5前4位进行散列。相应的，表1中二级索引的分区索引中的hashkey表示MD5值中任R位连续的数值。通过二级索引，具有相同后两位的appid对应的数据对象将被划分为一个数据分组集中处理，避免相同appid的数据对象集中

处理，引发底层系统性能瓶颈。进一步地，二级索引的分组索引key-range部分指向不同的三级索引，同时，key-range还可以按照规则创建时间和相同appid-prefix对待处理的数据对象进行排序。

[0076] 三级索引中存储有待处理的数据对象的存储信息。其中，filepath表示数据对象的存储路径信息，processtime表示数据对象的处理时间信息。其中，处理时间信息可以根据数据对象的修改时间来确定。本发明实施例中，可以按照二级索引的分区索引来组织三级索引的散列索引部分，这样，可以将数据对象的处理映射在不同的kv系统的表结构中。三级索引的分区索引部分可以按照文件修改时间的先后顺序排序，这样，可以按照用户需求先处理时间久的数据对象，再处理时间较新的数据对象。

[0077] 云服务器在接收到新的数据对象存储请求时，执行数据对象入库操作，即针对新的数据对象生成分组索引信息加入到索引结构中。首先根据数据对象对应的appid和存储桶标识(bucket)查询一级索引，以判断是否存在作用于该数据对象的数据处理规则，如果有，则可以按照以下方法生成该数据对象的二级索引和三级索引：根据用户标识中任两位相邻位置的数值生成二级索引中的散列索引，根据用户标识中任两位相邻位置的数值和根据数据对象的存储路径确定出的MD5值生成二级索引中的分区索引和三级索引中的散列索引；根据数据对象的修改时间信息确定数据对象的处理时间信息；并根据数据对象的处理时间信息和存储路径信息生成三级索引中的分区索引。在一个实施例中，可以取用户标识中的后两位作为二级索引中散列索引的appid_prefix；取MD5值的前四位作为二级索引中分区索引的hashkey等等，应该理解，上述实施方式仅用于举例说明，并不构成对本发明的限定。

[0078] 为了更好地理解本发明实施例，以下结合数据处理流程对本发明实施例的具体实施过程进行说明。本发明实施例中，以appid_prefix取用户标识中的后两位为例。由于appid_prefix取用户标识中的后两位，具体实施时，云服务器可以维护表2所示的9*9的bitmap数据结构，用于记录数据分组的处理进度，每一个坐标代表一个数据分组，处理完成的数据分组对应的坐标置为1，当bitmap中所有坐标值均为1时，则表示处理完成了所有的数据分组。

[0079] 表2

[0080]

	1	2	3	4	5	6	7	8	9
1									
2									
3									
4									
5									
6									
7									
8									
9									

[0081] 在当前处理周期开始时间到达时，可以随机选择一个起始坐标(i，j)，即当前时刻选择处理appid后两位为ij的所有数据对象，本发明实施例中，可以按照图3所示的流程实

施数据处理操作：

[0082] S31、根据二级索引的散列索引，获取坐标为(i，j)对应的所有分区索引。

[0083] 以起始坐标为(4,5)为例，即获取appid后两位为45的数据对象对应的所有分区索引，其中分区索引中进一步根据MD5值的前四位进行划分。

[0084] S32、选择一个未处理的分区索引。

[0085] 由于在二级索引的分区索引中，进一步根据MD5值的前4位对数据对象进行了划分，因此，本步骤中，可以选择一个分区索引进行处理。初始时，可以随机选择一个分区索引。

[0086] S33、以选择出的分区索引作为三级索引的散列索引，从对应的三级索引的分区索引中获取所有的数据对象存储信息。

[0087] 本步骤中，以步骤S32中选择出的分区索引作为三级索引的散列索引，并根据三级索引的散列索引从对应的分区索引中获取存储的数据对象存储信息。

[0088] S34、遍历获取的每一数据对象存储信息，根据存储信息中的时间处理信息，确定处理时间到达的目标数据对象。

[0089] S35、根据存储信息中的存储路径信息，并行处理所有的目标数据对象。

[0090] S36、判断是否处理完成所有的分区索引，如果是，则执行步骤S37，否则执行步骤S32。

[0091] S37、修改坐标(i，j)对应的坐标值为1。

[0092] S38、判断所有坐标值是否为1，如果是，流程结束，如果否，则执行步骤S39。

[0093] S39、判断j是否为最大值，如果是，执行步骤S310，否则，执行步骤S311。

[0094] 本发明实施例中，j的最大值即为9。

[0095] S310、修改i为i+1，修改j为j+1，并返回执行步骤S31。

[0096] S311、修改j为j+1，返回执行步骤S31。

[0097] 至此，完成了一个处理周期的数据对象处理，在每一处理周期，均可按照图3所示的流程对待处理的数据对象进行处理。

[0098] 本发明实施例提供的数据处理方法，根据分组索引信息将满足数据处理规则的数据对象划分为不同的数据分组进行处理，在每一数据处理周期，分别轮询每一分组，通过存储的存储路径信息能够快速定位到待处理的数据对象，并发处理同一数据分组中包含的所有数据对象，而且，通过特定的索引结构，降低了海量数据对象下特定规则的数据对象的选取，通过执行本发明实施例，可以快速并发处理大规模的数据对象处理。

[0099] 基于同一发明构思，本发明实施例中还提供了一种数据处理装置，由于上述装置解决问题的原理与数据处理方法相似，因此上述装置的实施可以参见方法的实施，重复之处不再赘述。

[0100] 如图4所示，其为本发明实施例提供的数据处理装置的结构示意图，包括：

[0101] 确定单元41，用于根据数据对象的分组索引信息，对属于同一个数据分组的各数据对象执行：根据各数据对象对应存储的处理条件，确定满足设定处理条件的目标数据对象；

[0102] 处理单元42，用于根据目标数据对象所对应存储的存储路径信息，并行处理属于同一个数据分组中的目标数据对象。

[0103] 可选地，所述分组索引信息根据所述数据对象对应的用户标识确定；或者所述分组索引信息根据所述数据对象的存储路径信息确定。

[0104] 可选地，所述分组索引信息由主索引信息和从索引信息组成。

[0105] 确定单元41，还可以用于根据所述数据对象对应的用户标识确定数据对象的主索引信息；针对根据所述主索引信息确定出的数据分组所包含的数据对象，进一步根据数据对象的存储路径信息确定数据对象的从索引信息。

[0106] 可选地，所述主索引信息为所述用户标识中任两位相邻位置的数值；以及所述从索引信息为根据所述数据对象的存储路径确定出的MD5值确定。

[0107] 可选地，所述分组索引信息中还包括数据处理规则和存储空间标识，以及所述数据对象为满足所述数据处理规则的数据对象，并且每一个存储空间标识对应一种数据处理规则。

[0108] 可选地，采用三级索引结构存储所述分组索引信息，其中，每一级索引由散列索引和分区索引组成，一级索引的散列索引中存储有业务标识，一级索引的分区索引中存储有所述数据对象对应的用户标识和存储空间标识，一级索引的值针对不同的存储空间标识存储其对应的数据处理规则；二级索引的散列索引中存储有根据所述用户标识中任两位相邻位置的数值生成所述分组索引信息，二级索引的分区索引中存储有根据所述数据对象的存储路径确定出的MD5值确定所述数据对象对应的分组索引信息；三级索引的散列索引中存储有根据所述数据对象的存储路径确定出的MD5值确定所述数据对象对应的分组索引信息；三级索引的分区索引中存储有所述数据对象对应的处理条件和存储路径信息。

[0109] 可选地，所述处理条件包括处理时间信息；以及

[0110] 所述装置，还包括：

[0111] 排序单元，用于根据所述处理时间信息，按照处理时间的先后顺序对三级索引的分区索引进行排序。

[0112] 为了描述的方便，以上各部分按照功能划分为各模块(或单元)分别描述。当然，在实施本发明时可以把各模块(或单元)的功能在同一个或多个软件或硬件中实现。

[0113] 在介绍了本发明示例性实施方式的数据处理方法和装置之后，接下来，介绍根据本发明的另一示例性实施方式的计算装置。

[0114] 所属技术领域的技术人员能够理解，本发明的各个方面可以实现为系统、方法或程序产品。因此，本发明的各个方面可以具体实现为以下形式，即：完全的硬件实施方式、完全的软件实施方式(包括固件、微代码等)，或硬件和软件方面结合的实施方式，这里可以统称为“电路”、“模块”或“系统”。

[0115] 在一些可能的实施方式中，根据本发明的计算装置可以至少包括至少一个处理器、以及至少一个存储器。其中，所述存储器存储有程序代码，当所述程序代码被所述处理器执行时，使得所述处理器执行本说明书上述描述的根据本发明各种示例性实施方式的数据处理方法中的步骤。例如，所述处理器可以执行如图2中所示的步骤S21、在每一处理周期，根据数据对象的分组索引信息，分别针对属于同一个数据分组的各数据对象执行：根据各数据对象对应存储的处理条件，确定该数据对象是否为满足设定处理条件的目标数据对象，和步骤S22、根据目标数据对象所对应存储的存储路径信息，并行处理属于同一个数据分组中的所有目标数据对象。

[0116] 下面参照图5来描述根据本发明的这种实施方式的计算装置50。图5显示的计算装置50仅仅是一个示例，不应对本发明实施例的功能和使用范围带来任何限制。

[0117] 如图5所示，计算装置50以通用计算设备的形式表现。计算装置50的组件可以包括但不限于：上述至少一个处理器51、上述至少一个存储器52、连接不同系统组件(包括存储器52和处理器51)的总线53。

[0118] 总线53表示几类总线结构中的一种或多种，包括存储器总线或者存储器控制器、外围总线、处理器或者使用多种总线结构中的任意总线结构的局域总线。

[0119] 存储器52可以包括易失性存储器形式的可读介质，例如随机存取存储器(RAM)521和/或高速缓存存储器522，还可以进一步包括只读存储器(ROM)523。

[0120] 存储器52还可以包括具有一组(至少一个)程序模块524的程序/实用工具525，这样的程序模块524包括但不限于：操作系统、一个或者多个应用程序、其它程序模块以及程序数据，这些示例中的每一个或某种组合中可能包括网络环境的实现。

[0121] 计算装置50也可以与一个或多个外部设备54(例如键盘、指向设备等)通信，还可与一个或者多个使得用户能与计算装置50交互的设备通信，和/或与使得该计算装置50能与一个或多个其它计算设备进行通信的任何设备(例如路由器、调制解调器等等)通信。这种通信可以通过输入/输出(I/O)接口55进行。并且，计算装置50还可以通过网络适配器56与一个或者多个网络(例如局域网(LAN)，广域网(WAN)和/或公共网络，例如因特网)通信。如图所示，网络适配器56通过总线53与用于计算装置50的其它模块通信。应当理解，尽管图中未示出，可以结合计算装置50使用其它硬件和/或软件模块，包括但不限于：微代码、设备驱动器、冗余处理器、外部磁盘驱动阵列、RAID系统、磁带驱动器以及数据备份存储系统等。

[0122] 在一些可能的实施方式中，本发明提供的数据处理方法的各个方面还可以实现为一种程序产品的形式，其包括程序代码，当所述程序产品在计算机设备上运行时，所述程序代码用于使所述计算机设备执行本说明书上述描述的根据本发明各种示例性实施方式的数据处理方法中的步骤，例如，所述计算机设备可以执行如图2中所示的步骤S21、在每一处理周期，根据数据对象的分组索引信息，分别针对属于同一个数据分组的各数据对象执行：根据各数据对象对应存储的处理条件，确定该数据对象是否为满足设定处理条件的目标数据对象，和步骤S22、根据目标数据对象所对应存储的存储路径信息，并行处理属于同一个数据分组中的所有目标数据对象。

[0123] 所述程序产品可以采用一个或多个可读介质的任意组合。可读介质可以是可读信号介质或者可读存储介质。可读存储介质例如可以是——但不限于——电、磁、光、电磁、红外线、或半导体的系统、装置或器件，或者任意以上的组合。可读存储介质的更具体的例子(非穷举的列表)包括：具有一个或多个导线的电连接、便携式盘、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)、光纤、便携式紧凑盘只读存储器(CD-ROM)、光存储器件、磁存储器件、或者上述的任意合适的组合。

[0124] 本发明的实施方式的用于数据处理的程序产品可以采用便携式紧凑盘只读存储器(CD-ROM)并包括程序代码，并可以在计算设备上运行。然而，本发明的程序产品不限于此，在本文件中，可读存储介质可以是任何包含或存储程序的有形介质，该程序可以被指令执行系统、装置或者器件使用或者与其结合使用。

[0125] 可读信号介质可以包括在基带中或者作为载波一部分传播的数据信号，其中承载

了可读程序代码。这种传播的数据信号可以采用多种形式，包括——但不限于——电磁信号、光信号或上述的任意合适的组合。可读信号介质还可以是可读存储介质以外的任何可读介质，该可读介质可以发送、传播或者传输用于由指令执行系统、装置或者器件使用或者与其结合使用的程序。

[0126] 可读介质上包含的程序代码可以用任何适当的介质传输，包括——但不限于——无线、有线、光缆、RF等等，或者上述的任意合适的组合。

[0127] 可以以一种或多种程序设计语言的任意组合来编写用于执行本发明操作的程序代码，所述程序设计语言包括面向对象的程序设计语言—诸如Java、C++等，还包括常规的过程式程序设计语言—诸如“C”语言或类似的程序设计语言。程序代码可以完全地在用户计算设备上执行、部分地在用户设备上执行、作为一个独立的软件包执行、部分在用户计算设备上部分在远程计算设备上执行、或者完全在远程计算设备或服务器上执行。在涉及远程计算设备的情形中，远程计算设备可以通过任意种类的网络——包括局域网(LAN)或广域网(WAN)—连接到用户计算设备，或者，可以连接到外部计算设备(例如利用因特网服务提供商来通过因特网连接)。

[0128] 应当注意，尽管在上文详细描述中提及了装置的若干单元或子单元，但是这种划分仅仅是示例性的并非强制性的。实际上，根据本发明的实施方式，上文描述的两个或更多单元的特征和功能可以在一个单元中具体化。反之，上文描述的一个单元的特征和功能可以进一步划分为由多个单元来具体化。

[0129] 此外，尽管在附图中以特定顺序描述了本发明方法的操作，但是，这并非要求或者暗示必须按照该特定顺序来执行这些操作，或是必须执行全部所示的操作才能实现期望的结果。附加地或备选地，可以省略某些步骤，将多个步骤合并为一个步骤执行，和/或将一个步骤分解为多个步骤执行。

[0130] 本领域内的技术人员应明白，本发明的实施例可提供为方法、系统、或计算机程序产品。因此，本发明可采用完全硬件实施例、完全软件实施例、或结合软件和硬件方面的实施例的形式。而且，本发明可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质(包括但不限于磁盘存储器、CD-ROM、光学存储器等)上实施的计算机程序产品的形式。

[0131] 本发明是参照根据本发明实施例的方法、设备(系统)、和计算机程序产品的流程图和/或方框图来描述的。应理解可由计算机程序指令实现流程图和/或方框图中的每一流程和/或方框、以及流程图和/或方框图中的流程和/或方框的结合。可提供这些计算机程序指令到通用计算机、专用计算机、嵌入式处理机或其他可编程数据处理设备的处理器以产生一个机器，使得通过计算机或其他可编程数据处理设备的处理器执行的指令产生用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的装置。

[0132] 这些计算机程序指令也可存储在能引导计算机或其他可编程数据处理设备以特定方式工作的计算机可读存储器中，使得存储在该计算机可读存储器中的指令产生包括指令装置的制造品，该指令装置实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能。

[0133] 这些计算机程序指令也可装载到计算机或其他可编程数据处理设备上，使得在计算机或其他可编程设备上执行一系列操作步骤以产生计算机实现的处理，从而在计算机或

其他可编程设备上执行的指令提供用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的步骤。

[0134] 尽管已描述了本发明的优选实施例，但本领域内的技术人员一旦得知了基本创造性概念，则可对这些实施例做出另外的变更和修改。所以，所附权利要求意欲解释为包括优选实施例以及落入本发明范围的所有变更和修改。

[0135] 显然，本领域的技术人员可以对本发明进行各种改动和变型而不脱离本发明的精神和范围。这样，倘若本发明的这些修改和变型属于本发明权利要求及其等同技术的范围之内，则本发明也意图包含这些改动和变型在内。

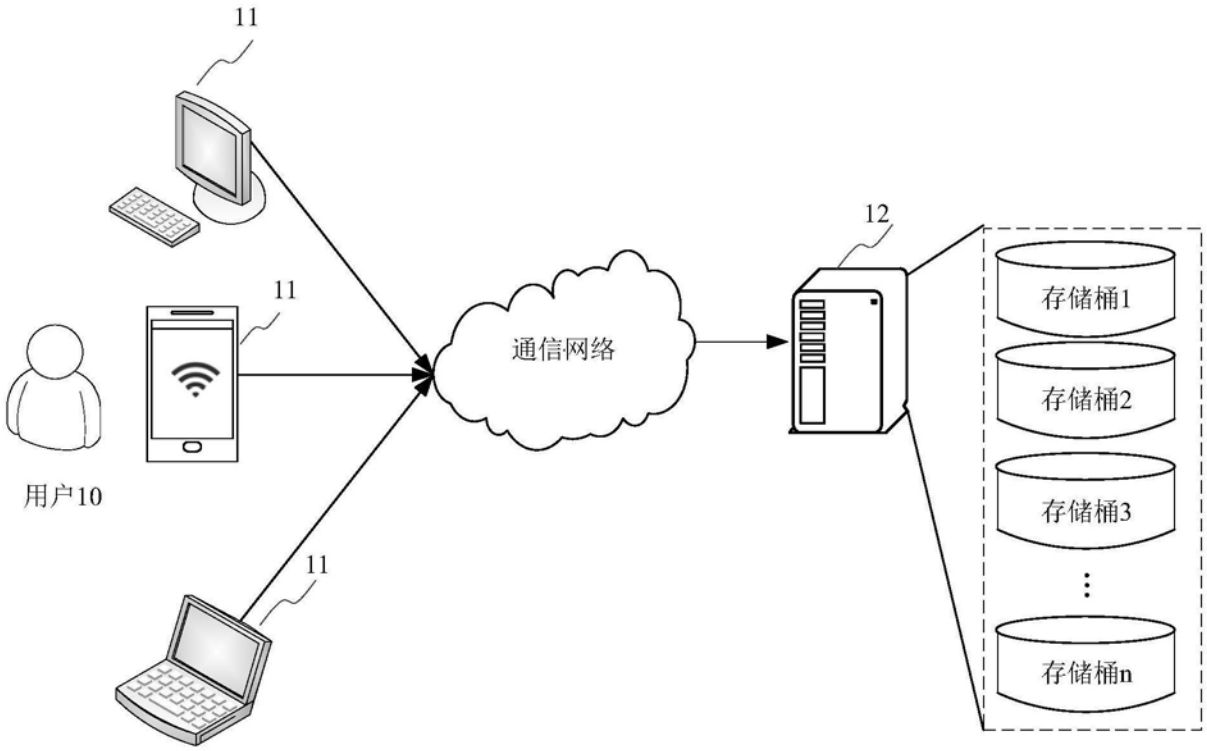

11
11
11
用户10
通信网络
12
存储桶1
存储桶2
存储桶3
存储桶n

图1

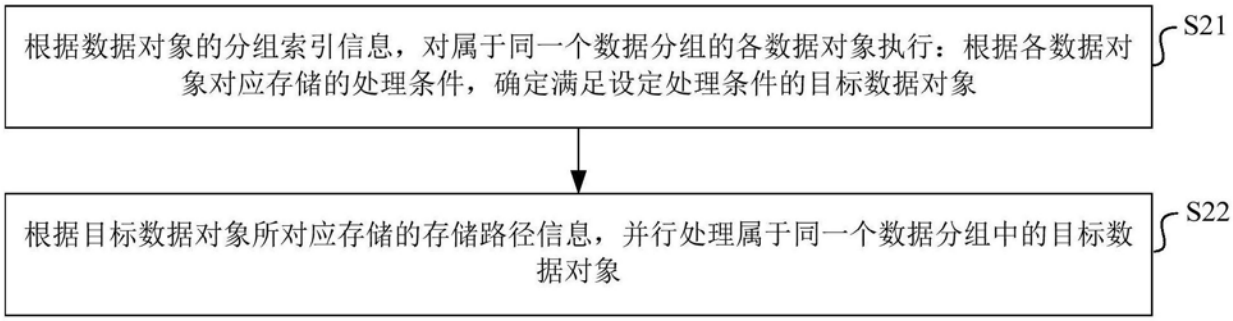

根据数据对象的分组索引信息，对属于同一个数据分组的各数据对象执行：根据各数据对象对应存储的处理条件，确定满足设定处理条件的目标数据对象
S21
根据目标数据对象所对应存储的存储路径信息，并行处理属于同一个数据分组中的目标数据对象
S22

图2

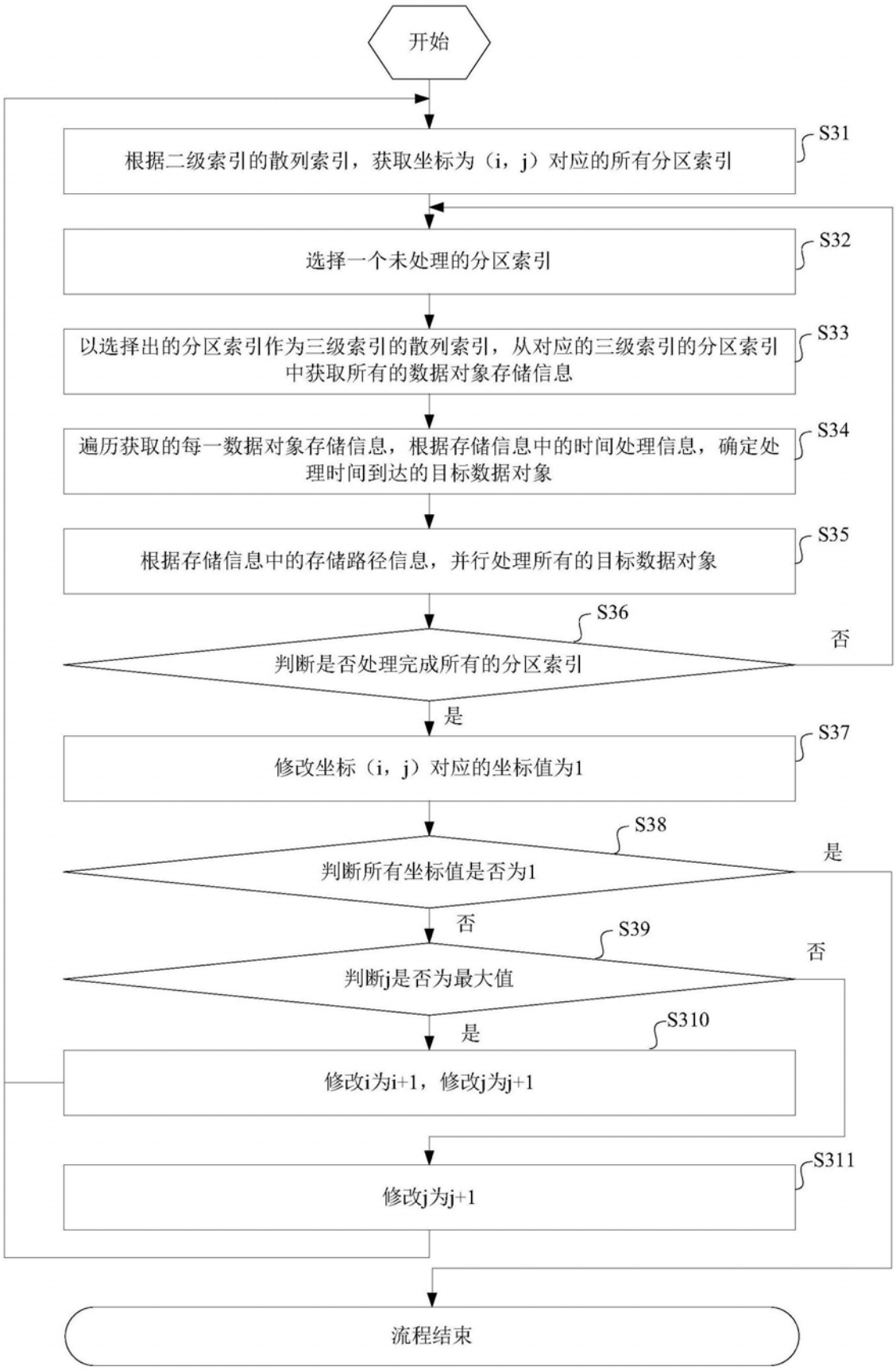

开始
根据二级索引的散列索引，获取坐标为（i，j）对应的所有分区索引
S31
选择一个未处理的分区索引
S32
以选择出的分区索引作为三级索引的散列索引，从对应的三级索引的分区索引中获取所有的数据对象存储信息
S33
遍历获取的每一数据对象存储信息，根据存储信息中的时间处理信息，确定处理时间到达的目标数据对象
S34
根据存储信息中的存储路径信息，并行处理所有的目标数据对象
S35
判断是否处理完成所有的分区索引
S36
否
是
修改坐标（i，j）对应的坐标值为1
S37
判断所有坐标值是否为1
S38
是
否
判断j是否为最大值
S39
否
是
修改i为i+1，修改j为j+1
S310
修改j为j+1
S311
流程结束

图3

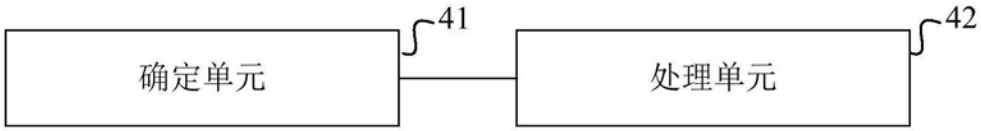

41
确定单元
42
处理单元

图4

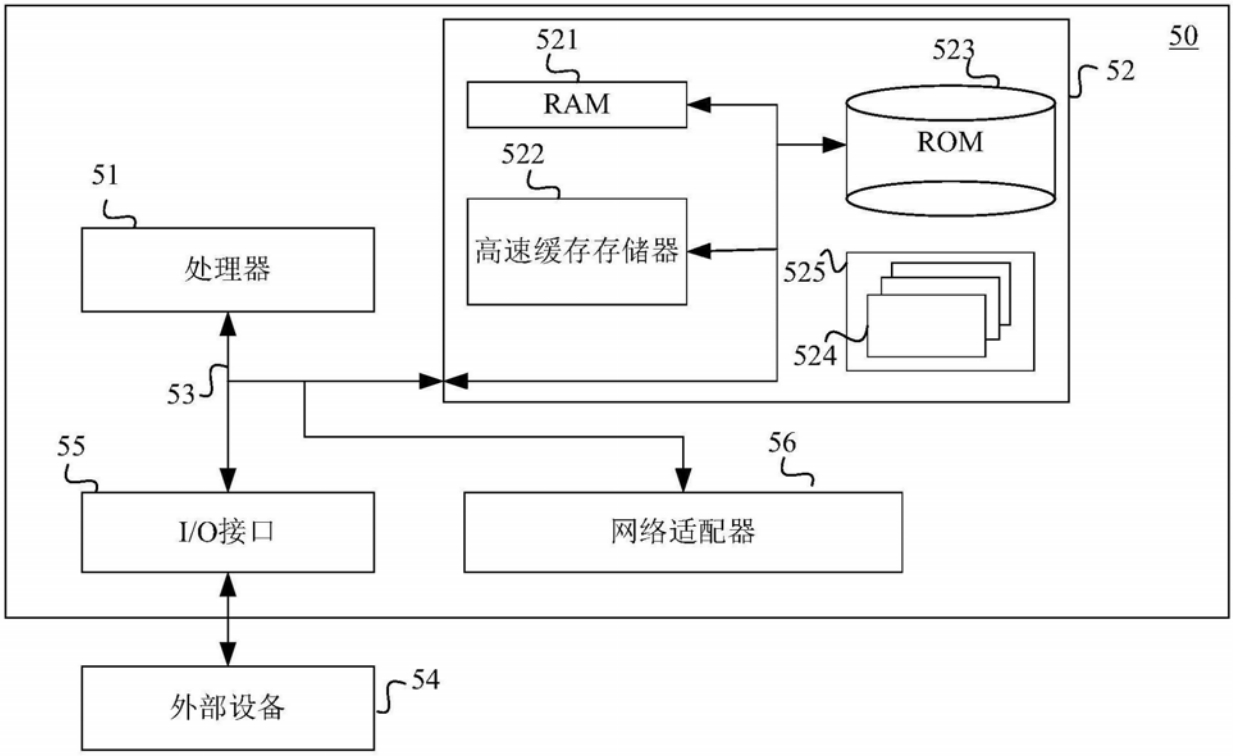

50
521
RAM
523
ROM
52
522
高速缓存存储器
525
524
51
处理器
53
55
I/O接口
56
网络适配器
外部设备
54

图5