(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 103310000 B
(45)授权公告日 2017.06.16

(21)申请号 201310256313.9

(22)申请日 2013.06.25

(65)同一申请的已公布的文献号
申请公布号 CN 103310000 A

(43)申请公布日 2013.09.18

(73)专利权人 曙光信息产业(北京)有限公司
地址 100193 北京市海淀区东北旺西路8号院36号楼

(72)发明人 杨浩 马照云 马振杰 苗艳超 刘新春 邵宗有

(74)专利代理机构 北京德恒律治知识产权代理有限公司 11409
代理人 章社杲 孙征

(51)Int.Cl.
G06F 17/30(2006.01)

(56)对比文件
CN 101408880 A,2009.04.15,
US 2007/0239949 A1,2007.10.11,
CN 103067461 A,2013.04.24,

审查员 孔昕

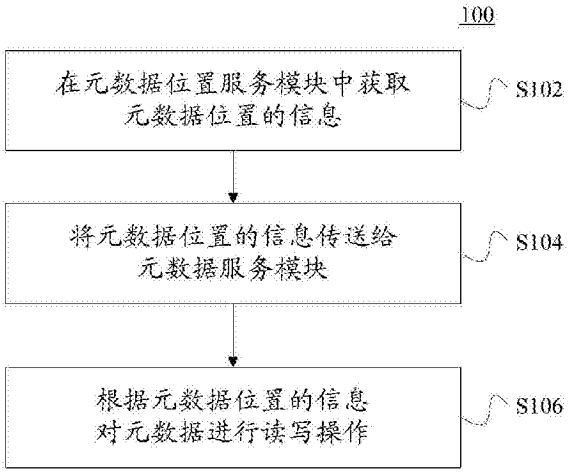
权利要求书2页 说明书6页 附图3页

(54)发明名称

元数据管理方法

(57)摘要

本发明提供了一种元数据管理方法，包括：在元数据位置服务模块中获取元数据位置的信息；将元数据位置的信息传送给元数据服务模块；以及根据元数据位置的信息对元数据进行读写操作。本发明将元数据的管理划分为元数据位置管理层和元数据管理层，通过元数据位置管理层统一管理用户访问的元数据位置信息，然后元数据管理层根据元数据位置的信息将用户的访问请求分散到多个服务器上去执行，从而能够克服由于多个用户同时访问元数据存储系统所导致的系统响应时间长、反应慢、访问效率低等缺陷，因此，能够承受多个用户同时访问该存储系统所带来的访问压力，从而提高了系统性能。



CN 103310000 B

1.一种元数据管理方法,其特征在于,包括:

在元数据位置服务模块中获取元数据位置的信息;

将所述元数据位置的信息传送给元数据服务模块;以及

根据所述元数据位置的信息对元数据进行读写操作;

在实施所述元数据管理方法的步骤之前,进一步包括:通过客户端向所述元数据位置服务模块发送元数据操作请求,其中,所述元数据操作请求包括:查询请求和存储请求;

其中,所述元数据位置为全局唯一标识ID,在所述元数据管理方法的步骤开始之前,进一步包括:

建立所有全局唯一标识ID和元数据存储位置之间的一致性哈希映射关系,并将所述元数据存储位置作为元数据存储位置集合;

将所述元数据存储位置集合划分为M个子集,其中,M为大于N的整数并且N为元数据存储节点的个数;

将所述M个子集分配给N个元数据存储节点;以及

根据所述一致性哈希映射关系,将具有所述全局唯一标识ID的元数据存储在与所述M个子集相对应的所述N个元数据存储节点上;

当所述N个元数据存储节点的存储空间不足时,在线添加扩展元数据存储节点,并且将存储在所述N个元数据存储节点上的M个子集的一部分转存至所述扩展元数据存储节点上;

其次,将所述元数据位置服务模块中的存储请求和访问请求进行分离。

2.根据权利要求1所述的元数据管理方法,其特征在于,所述元数据位置服务模块位于一个或多个存储节点组上,每个存储节点组都包括一个主存储节点和一个从存储节点,并且所述主存储节点和所述从存储节点互为镜像存储节点。

3.根据权利要求2所述的元数据管理方法,其特征在于,当所述一个或多个存储节点组的存储空间不足时,以存储节点组为单位在线添加扩展存储节点组,并且将后续要写入的所述元数据的所述元数据位置的信息存储在所述扩展存储节点组上。

4.根据权利要求1所述的元数据管理方法,其特征在于,当所述元数据操作请求为所述查询请求时,

根据所述查询请求,在所述元数据位置服务模块中查找到所述元数据的全局唯一标识ID;

将所述全局唯一标识ID传送给所述元数据服务模块;以及

根据所述全局唯一标识ID的一致性哈希映射查找到存储所述元数据的元数据存储节点,从所述元数据存储节点读取所述元数据并反馈给所述客户端。

5.根据权利要求1所述的元数据管理方法,其特征在于,当所述元数据操作请求为所述存储请求时,

根据所述存储请求,由所述元数据位置服务模块提供所述元数据的全局唯一标识ID;

将所述全局唯一标识ID和所述元数据传送给所述元数据服务模块;以及

根据所述全局唯一标识ID的所述一致性哈希映射查找到要存储所述元数据的元数据存储节点,并将所述元数据存储在所述元数据存储节点中。

6.根据权利要求1至5中的任一项所述的元数据管理方法,其特征在于,所述元数据服务模块用于管理包括文件的基本属性、用户数据存储位置信息以及文件扩展属性的元数

据。

7.根据权利要求1至5中的任一项所述的元数据管理方法，其特征在于，所述元数据位置服务模块用于管理所述元数据位置信息和文件系统的目录树结构。

元数据管理方法

技术领域

[0001] 本发明一般地涉及计算机技术领域，更具体地来说，涉及元数据管理方法。

背景技术

[0002] 随着科技的日异月新，文件数据的指数性增长给存储系统带来了巨大的挑战。与传统的SAN结构存储比较起来，分布式文件系统具有价格低廉、可扩展性强、性能优越等特性。作为通用的分布式文件系统，元数据的管理是需要重点关注的，因为元数据是管理数据的数据。根据访问特性，应用可分为两种形式：元数据非密集型和元数据密集型。元数据非密集型应用主要是大文件的访存，如视频监控、虚拟机应用等，这类应用集中于数据的访存，元数据访问的比例相对较低，因而对于元数据管理的压力较小；元数据密集型应用集中于小文件的访存，如数字图书馆、网上商城等，这类应用操作的文件量巨大，元数据的访问压力较大。当管理海量的小文件时，分布式文件系统需要面临着性能和存储容量的双重考验。

[0003] 在现有技术中，对于海量小文件的管理均存在一定的局限性。Lustre、GoogleFS、HDFS采用单一元数据服务节点架构，无法进行扩展，因此不论是在性能还是存储容量上都大大受限，不能很好地满足海量小文件管理的需求。GPFS采用多元数据，但是元数据服务器之间采用分布式锁来维持一致性，因此当元数据服务节点的数量较多时，分布式锁的开销会激增，对性能照成较大的负面影响。Panasas的存储系统采用多元数据服务器架构，但是每个元数据节点管理的命名空间是独立的，不能实现真正意义上的全局命名空间。Ceph采用动态子树的方式来在多个元数据节点之间平衡元数据访问负载，这种方式可以实现访问负载的均衡，但其不足之处在于元数据节点不能实现存储容量的动态扩展。因此，现有技术中的分布式文件系统不能同时满足巨大的应用操作的文件量和较大的元数据的访问压力方面的要求。

发明内容

[0004] 针对现有技术中的存在分布式文件系统不能同时满足巨大的应用操作的文件量和较大的元数据的访问压力方面的要求的缺陷，本发明提供了能够解决上述缺陷的元数据管理方法。

[0005] 根据本发明的一方面，本发明提供了一种元数据管理方法，包括：在元数据位置服务模块中获取元数据位置的信息；将元数据位置的信息传送给元数据服务模块；以及根据元数据位置的信息对元数据进行读写操作。

[0006] 优选地，元数据位置为全局唯一标识ID，在元数据管理方法的步骤开始之前，进一步包括：建立所有全局唯一标识ID和元数据存储位置之间的一致性哈希映射关系，并将元数据存储位置作为元数据存储位置集合；将元数据存储位置集合划分为M个子集，其中，M为大于N的整数并且N为元数据存储节点的个数；将M个子集分配给N个元数据存储节点；以及根据一致性哈希映射关系，将具有全局唯一标识ID的元数据存储在与M个子集相对应的N个

元数据存储节点上。

[0007] 优选地，当N个元数据存储节点的存储空间不足时，在线添加扩展元数据存储节点，并且将存储在N个元数据存储节点上的M个子集的一部分转存至扩展元数据存储节点上。

[0008] 优选地，元数据位置服务模块位于一个或多个存储节点组上，每个存储节点组都包括一个主存储节点和一个从存储节点，并且主存储节点和从存储节点互为镜像存储节点。

[0009] 优选地，当一个或多个存储节点组的存储空间不足时，以存储节点组为单位在线添加扩展存储节点组，并且将后续要写入的元数据的元数据位置的信息存储在扩展存储节点组上。

[0010] 优选地，在实施元数据管理方法的步骤之前，进一步包括：通过客户端向元数据位置服务模块发送元数据操作请求，其中，元数据操作请求包括：查询请求和存储请求。

[0011] 优选地，当元数据操作请求为查询请求时，根据查询请求，在元数据位置服务模块中查找到元数据的全局唯一标识ID；将全局唯一标识ID传送给元数据服务模块；以及根据全局唯一标识ID的一致性哈希映射查找到存储元数据的元数据存储节点，从元数据存储节点读取元数据并反馈给客户端。

[0012] 优选地，当元数据操作请求为存储请求时，根据存储请求，由元数据位置服务模块提供元数据的全局唯一标识ID；将全局唯一标识ID和元数据传送给元数据服务模块；以及根据全局唯一标识ID的一致性哈希映射查找到要存储元数据的元数据存储节点，并将元数据存储在元数据存储节点中。

[0013] 优选地，元数据服务模块用于管理包括文件的基本属性、用户数据存储位置信息以及文件扩展属性的元数据。

[0014] 优选地，元数据位置服务模块用于管理元数据位置信息和文件系统的目录树结构。

[0015] 利用本发明的技术方案能够克服现有技术的缺陷，通过将文件系统的元数据管理划分为元数据位置管理层和元数据管理层，这样能够解决多个用户同时访问海量小文件的压力，而且这种元数据管理方法便于在线扩展元数据位置存储节点和元数据存储节点，从而便于动态增大存储容量。因此能够满足巨大的应用操作的文件量和较大的元数据的访问压力方面的要求。

[0016] 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

[0017] 以下结合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。在附图中：

[0018] 图1为根据本发明的实施例的元数据管理方法的整体流程图；

[0019] 图2为根据本发明的实施例的元数据管理方法的存储方法的具体实例的示图；

[0020] 图3为根据本发明的实施例在实施元数据管理方法之前，建立元数据存储结构的

方法的步骤的流程图；

[0021] 图4为本发明的可选实施例的建立的部分元数据存储节点的示图；

[0022] 图5为根据本发明的实施例的元数据管理方法的查询方法的具体流程的示图；以及

[0023] 图6为根据本发明的实施例的元数据管理方法的存储方法的具体流程图。

具体实施方式

[0024] 以下结合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。

[0025] 图1为根据本发明的实施例的元数据管理方法的整体流程图。参考图1，元数据管理方法100包括以下步骤。在步骤S102中，在元数据位置服务模块中获取元数据位置的信息。具体地，客户端接收到来自用户的数据请求以后，首先将接收到的数据请求传送给元数据位置服务模块，并且在元数据位置服务模块中获取元数据位置信息。在步骤S104中，将元数据位置的信息传送给元数据服务模块。具体地，元数据位置服务模块将获得的元数据位置的信息传送给元数据服务模块。在步骤S106中，根据元数据位置的信息对元数据进行读写操作。具体地，元数据服务模块在接收到元数据位置的信息以后，根据该元数据位置的信息对该元数据进行读操作或写操作。

[0026] 本发明的实施例的元数据管理方法将元数据的管理划分为元数据位置管理层和元数据管理层，通过元数据位置管理层统一管理用户访问的元数据位置信息，然后元数据管理层根据元数据位置的信息将用户的访问请求分散到多个服务器上去执行，从而能够克服由于多个用户同时访问元数据存储系统所导致的系统响应时间长、反应慢、访问效率低等缺陷，因此，能够承受多个用户同时访问该存储系统所带来的访问压力，从而提高了系统性能。

[0027] 本发明中对文件系统元数据管理采取分层管理，共分为两层：第一层为元数据位置管理层，第二层为元数据管理层。元数据位置管理层用于管理文件元数据所在的元数据存储节点位置信息和文件系统的目录树结构。具体地，元数据位置管理层(即，元数据位置服务模块)向用户的数据请求接收层提供各个文件的具体元数据存储位置信息以及传统的目录树结构。元数据服务模块用于管理包括文件的基本属性、用户数据存储位置信息以及文件扩展属性的元数据。具体地，元数据管理层用于管理文件的基本属性(如大小、修改时间)、用户数据存储位置信息、文件扩展属性(如权限配置)等。在海量小文件存储系统中，由于需要管理海量文件，元数据量较多，因此元数据管理层的元数据信息存储需求较元数据位置管理层的元数据位置信息存储需求大。图2为根据本发明的实施例的元数据管理方法的存储方法的具体实例的示图。应用访问元数据的流程如下：

[0028] (1)应用将向客户端发起请求；

[0029] (2)向元数据位置服务模块查询元数据的存储位置；

[0030] (3)元数据位置服务模块将请求转发给元数据服务模块；

[0031] (4)元数据服务模块读取元数据，返回给客户端；以及

[0032] (5)客户端向应用返回请求处理结果。

[0033] 因此，利用该元数据管理方法解决了分布式文件系统中元数据性能和存储容量所

面临的双重压力问题。

[0034] 在元数据管理方法的步骤开始之前，首先需要建立能够实现参考图1所述的元数据管理方法的元数据存储结构(即，分布式文件系统)。图3为根据本发明的实施例在实施元数据管理方法之前，建立元数据存储结构的方法的步骤的流程图。参考图3，建立元数据存储结构的方法300包括以下具体步骤。在步骤S302中，建立所有全局唯一标识ID和元数据存储位置之间的一致性哈希映射关系，并将元数据存储位置作为元数据存储位置集合。其中元数据位置为全局唯一标识ID。具体地，将该元数据存储系统中的所有的元数据位置进行统一编号，从而便于在整个元数据存储系统中进行全局查找。具体地，通过一致性哈希映射建立全局唯一标识ID和实际的元数据存储位置之间的映射关系，并且将实际的元数据存储位置作为元数据存储位置集合。

[0035] 在步骤S304中，将元数据存储位置集合划分为M个子集，其中，M为大于N的整数并且N为元数据存储节点的个数。具体地，对实际的元数据存储位置集合进行划分，当存储元数据的元数据存储节点具有N个时，将元数据存储位置的集合划分为M个，其中M远远大于N，从而能够轻松地实现元数据存储节点的在线扩展。

[0036] 在步骤S306中，将M个子集分配给N个元数据存储节点。对元数据存储位置的M个子集进行分配，在建立存储结构时，将M个子集分配给N个元数据存储节点，当元数据存储节点的存储容量不足时，可以将具有多个子集的存储节点中的一个或多个子集分配给扩展的元数据存储节点，从而使元数据存储节点能够进行在线扩展。

[0037] 在步骤S308中，根据一致性哈希映射关系，将具有全局唯一标识ID的元数据存储在与M个子集相对应的N个元数据存储节点上。具体地，每个元数据都会具有一个全局唯一标识ID，将具有全局唯一标识ID的相应元数据存储在分配有相应的元数据存储位置子集的存储节点上，从而建立能够通过本发明实施例的元数据管理方法所管理的存储结构(即，分布式文件系统)。

[0038] 在通过建立元数据存储结构的方法300所建立的元数据存储结构中，元数据位置管理层(即，元数据位置服务模块)采用树结构管理，对应传统文件系统的目录项管理。该层的存储量不大，通常TB级的存储容量就可以满足存储容量的要求。由于元数据存储位置的访问需求较大，且在常规应用模式下，以查询访问为主。优选地，采取存储和访问分离的结构，即，存储节点数目和提供定位查询服务的访问节点可以不同，以避免元数据位置服务模块所提供的服务成为整个存储系统的瓶颈。例如，由于该分布式文件系统，主要是提供访问服务，所以可以将更多的存储节点分配给访问请求，而分配给存储请求的存储节点较少。当客户端接收到应用请求时，将元数据位置查询任务下发给元数据位置服务模块进行处理，元数据位置服务模块通常操作内存缓存，在必要时，元数据位置服务模块进行磁盘访问操作。

[0039] 在元数据管理层(即，元数据服务模块)中，仅涉及到文件级别元数据的管理，目录级别的元数据管理由其上层(即，元数据位置管理层)进行管理。对于存储系统中的每一个文件，都具有一个全局唯一的标识(ID)，该ID是在文件生成时，由元数据位置服务模块负责分配的。该元数据服务模块管理的元数据信息包括文件的大小、时间信息、数据分布信息等，由于此类信息需要的存储空间较大，因此需要多台存储节点共同承担，同时采用节点内部RAID6和节点间副本技术，来实现元数据冗余。在可选实施例中，在元数据信息存储时，将

全局ID分为若干区间，每个区间的元数据信息存储在一个存储节点，以实现存储空间的均衡(参考图4)。

[0040] 以下将对元数据位置存储节点和元数据存储节点进行描述。

[0041] 当N个元数据存储节点的存储空间不足时，在线添加扩展元数据存储节点，并且将存储在N个元数据存储节点上的M个子集的一部分转存至扩展元数据存储节点上。具体地，随着存储文件的增加，在该存储结构的存储空间不足时，可以在线添加扩展元数据存储节点，并且将存储在N个元数据存储节点上的M个子集的一部分子集转存至新添加的元数据存储节点上，从而，在不影响该分布式文件系统的工作的情况下，也就是不必关闭系统就可以扩展存储空间的容量。

[0042] 元数据位置服务模块位于一个或多个存储节点组上，每个存储节点组都包括一个主存储节点和一个从存储节点，并且主存储节点和从存储节点互为镜像存储节点。具体地，元数据位置服务模块位于一个或多个存储节点组(即，元数据位置存储节点对)上。当一个或多个存储节点组的存储空间不足时，以存储节点组为单位在线添加扩展存储节点组，并且将后续要写入的元数据的元数据位置的信息存储在扩展存储节点组上。具体地，当该分布式文件系统的一个或多个存储节点组的存储空间不足时，可以以存储节点组为单位在线添加扩展节点组，并且将以后的数据请求要写入的元件数据的元数据位置的信息存储在扩展节点组上来实现元数据位置存储节点对的扩容。

[0043] 以下将分别参考图5和图6描述元数据管理方法中的查询请求和存储请求的具体步骤。

[0044] 在实施元数据管理方法的步骤之前，进一步包括：通过客户端向元数据位置服务模块发送元数据操作请求，其中，元数据操作请求包括：查询请求和存储请求。

[0045] 图5为根据本发明的实施例的元数据管理方法的查询方法的具体流程的示图。参考图5，元数据管理方法的查询方法500包括以下具体步骤。当元数据操作请求为查询请求时，在步骤S502中，根据查询请求，在元数据位置服务模块中查找到元数据的全局唯一标识ID。具体地，当数据请求为查询请求时，根据查询请求的信息，在元数据位置服务模块中查找到元数据的全局唯一标识ID。在步骤S504中，将全局唯一标识ID传送给元数据服务模块。具体地，元数据位置服务模块在查找到元数据的全局唯一标识ID以后，将该全局唯一标识ID传送给元数据服务模块。在步骤S506中，根据所述全局唯一标识ID的一致性哈希映射查找到存储所述元数据的元数据存储节点，从所述元数据存储节点读取所述元数据并反馈给所述客户端。具体地，在元数据服务模块中，对全局唯一标识ID进行一致性哈希映射，因此，根据全局唯一标识ID的一致性哈希映射可以确定全局唯一标识ID位于M个子集中的哪个子集中，进而可以确定存储元数据的元数据存储节点，然后，元数据服务模块从元数据存储节点读取元数据并将读取的元数据通过客户端反馈给用户。

[0046] 图6为根据本发明的实施例的元数据管理方法的存储方法的具体流程图。参考图6，元数据管理方法的存储方法600包括以下具体步骤。当元数据操作请求为存储请求时，在步骤S602中，根据存储请求，由元数据位置服务模块提供元数据的全局唯一标识ID。具体地，当元数据操作请求为存储请求时，元数据位置服务模块为该存储请求的要存储的元数据分配一个全局唯一标识ID。在步骤S604中，将全局唯一标识ID和元数据传送给元数据服务模块。具体地，元数据位置服务模块在分配完全局唯一标识ID以后，将分配的全局唯一标

识ID和要存储的元数据传送给元数据服务模块。在步骤S606中，根据全局唯一标识ID的一致性哈希映射查找到要存储元数据的元数据存储节点，并将元数据存储在元数据存储节点中。具体地，在元数据服务模块中，对分配的全局唯一标识ID进行一致性哈希映射，因此，根据全局唯一标识ID的一致性哈希映射可以确定全局唯一标识ID位于M个子集中的哪个子集中，进而可以确定要存储的元数据的元数据存储节点，然后，元数据服务模块将要存储的元数据存储在确定的元数据存储节点上，然后将存储成功的信息和全局唯一标识ID通过客户端反馈给用户。

[0047] 利用本发明实施例的元数据管理方法，能够获得以下技术效果：首先，在建立该分布式文件系统时，将元数据的管理划分为元数据位置管理层和元数据管理层，通过元数据位置管理层统一管理用户访问的元数据位置信息，然后元数据管理层根据元数据位置的信息将用户的访问请求分散到多个服务器上去执行，从而能够克服由于多个用户同时访问元数据存储系统所导致的系统响应时间长、反应慢、访问效率低等缺陷，因此，能够承受多个用户同时访问该存储系统所带来的访问压力，从而提高了系统性能。其次，该分布式文件系统的结构允许在线扩展元数据位置存储节点和元数据存储节点，从而解决了海量文件不断增加所引起的存储空间不足的问题。此外，可以将元数据位置服务模块中的存储请求和访问请求进行分离，从而避免了元数据位置服务模块成为整个存储系统的瓶颈。

[0048] 以上仅为本发明的优选实施例而已，并不用于限制本发明，对于本领域的技术人员来说，本发明可以有各种更改和变化。凡在本发明的精神和原则之内，所作的任何修改、等同替换、改进等，均应包含在本发明的保护范围之内。

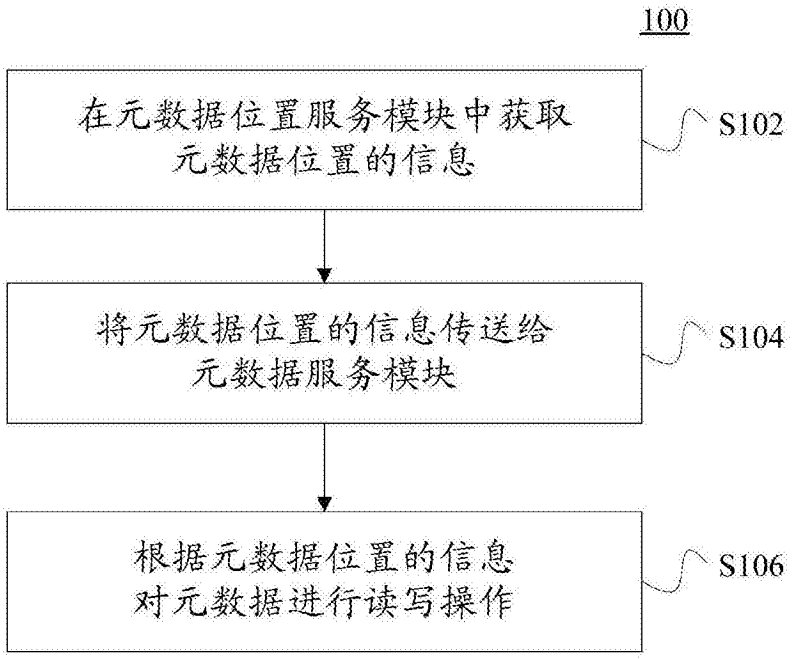

100
在元数据位置服务模块中获取
元数据位置的信息
S102
将元数据位置的信息传送给
元数据服务模块
S104
根据元数据位置的信息
对元数据进行读写操作
S106

图1

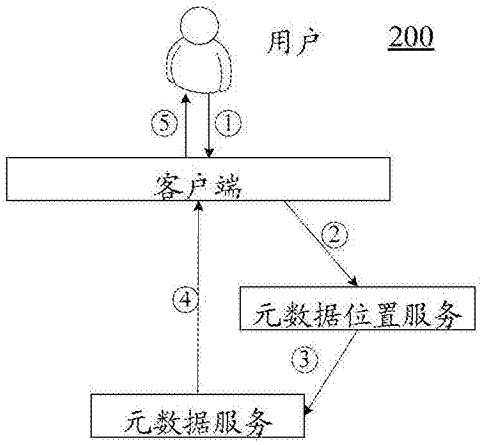

用户
200
⑤
①
客户端
②
④
元数据位置服务
③
元数据服务

图2

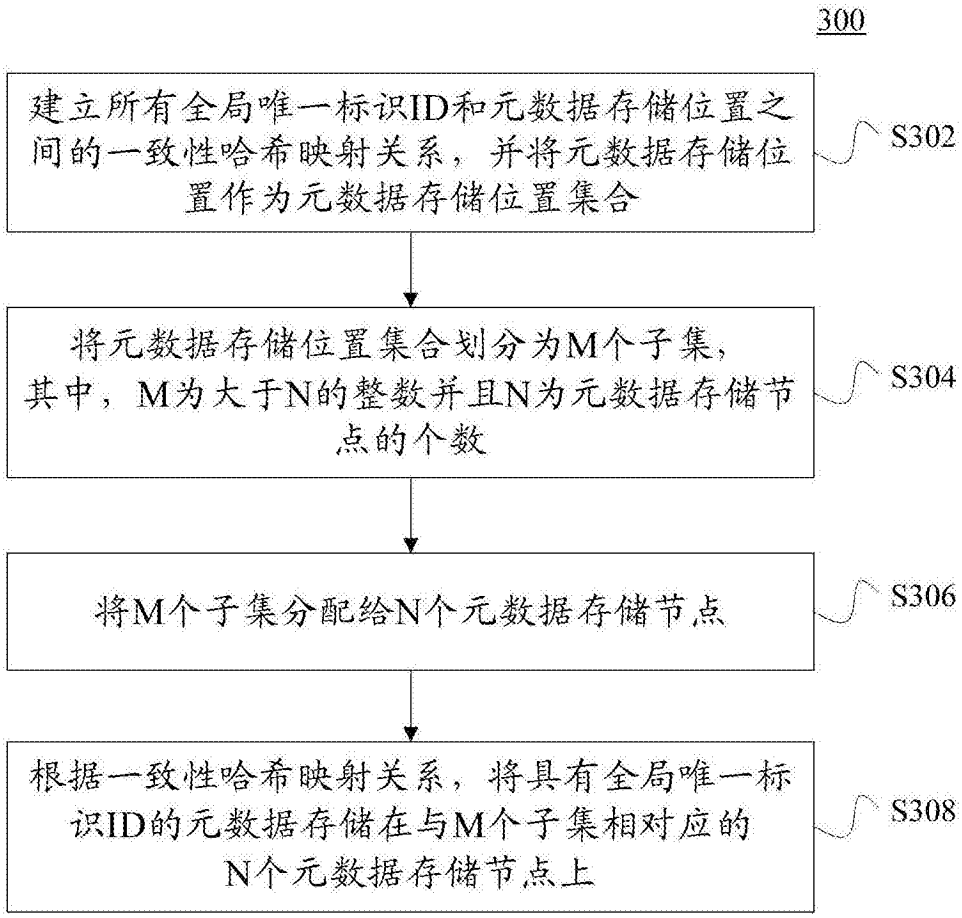

300
建立所有全局唯一标识ID和元数据存储位置之间的一致性哈希映射关系，并将元数据存储位置作为元数据存储位置集合
S302
将元数据存储位置集合划分为M个子集，其中，M为大于N的整数并且N为元数据存储节点的个数
S304
将M个子集分配给N个元数据存储节点
S306
根据一致性哈希映射关系，将具有全局唯一标识ID的元数据存储在与M个子集相对应的N个元数据存储节点上
S308

图3

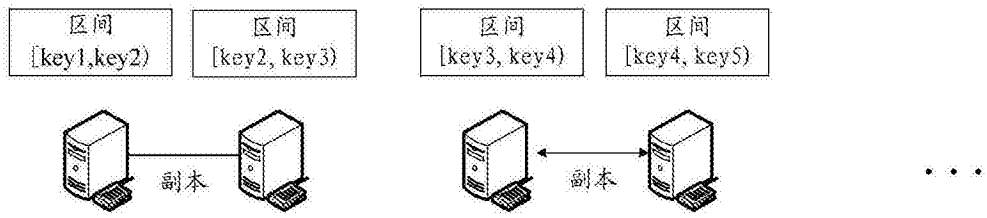

区间
[key1,key2)
区间
[key2, key3)
区间
[key3, key4)
区间
[key4, key5)
副本
副本
. . .

图4

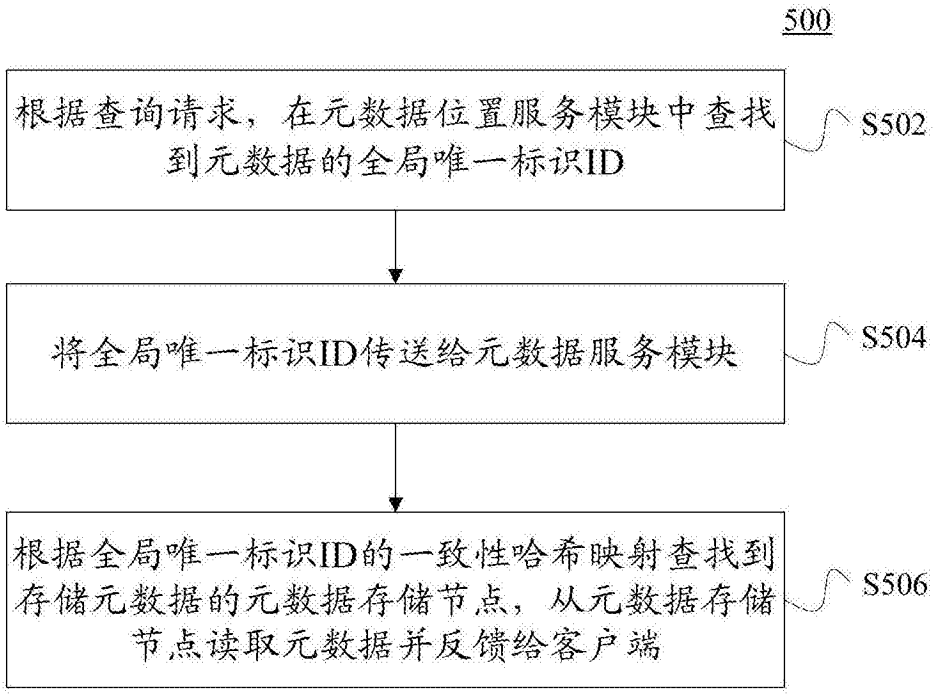

500
根据查询请求，在元数据位置服务模块中查找到元数据的全局唯一标识ID
S502
将全局唯一标识ID传送给元数据服务模块
S504
根据全局唯一标识ID的一致性哈希映射查找到存储元数据的元数据存储节点，从元数据存储节点读取元数据并反馈给客户端
S506

图5

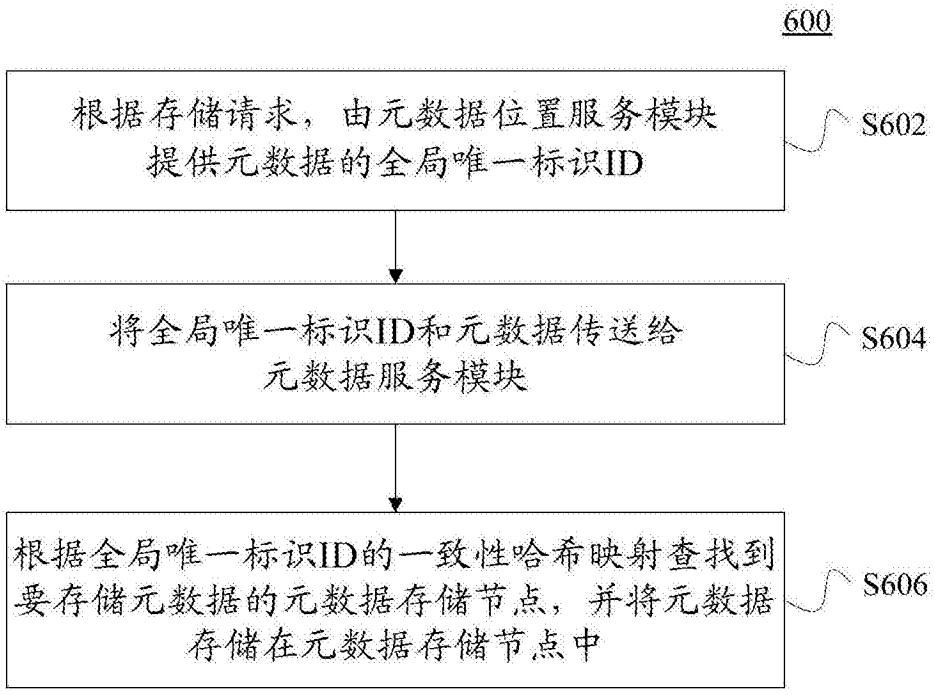

600
根据存储请求，由元数据位置服务模块提供元数据的全局唯一标识ID
S602
将全局唯一标识ID和元数据传送给元数据服务模块
S604
根据全局唯一标识ID的一致性哈希映射查找到要存储元数据的元数据存储节点，并将元数据存储在元数据存储节点中
S606

图6