(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 102833331 B
(45)授权公告日 2015.06.10

(21)申请号 201210299624.9

(22)申请日 2012.08.21

(73)专利权人 北京邦诺存储科技有限公司
地址 100085 北京市海淀区上地信息路 12 号中关村发展大厦 B 座 101 室

(72)发明人 严杰 熊晖 周娟娟

(74)专利代理机构 北京聿宏知识产权代理有限公司 11372
代理人 钟日红 王建军

(51)Int.Cl.
H04L 29/08(2006.01)
G06F 17/30(2006.01)

(56)对比文件
CN 102243660 A,2011.11.16,
WO 2012024801 A1,2012.03.01,
CN 102158546 A,2011.08.17,
李锐等.《空间数据存储对象的元数据可伸缩性管理》.《计算机应用研究》.2011,第 28 卷(第 12 期),4567-4571.

审查员 胡延

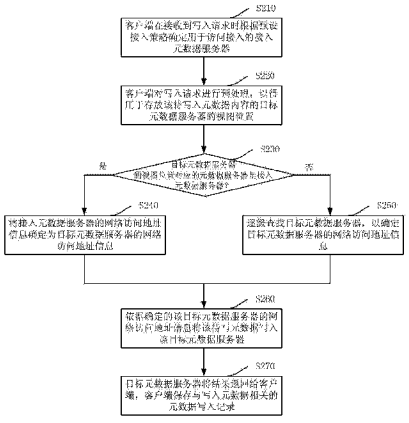
权利要求书3页 说明书9页 附图3页

(54)发明名称

云存储系统及其元数据写入方法、元数据读取方法

(57)摘要

本发明公开了一种云存储系统及其元数据写入方法、元数据读取方法。该系统包括:包括元数据服务器集群、视图管理模块和客户端。客户端向用户提供访问云存储系统的接口并解析用户要访问的元数据服务器的视图位置;视图管理模块存储全部元数据视图信息,根据全部元数据视图信息构建各元数据服务器之间的逻辑关联关系,并向每个元数据服务器分发与元数据服务器自身相关联的元数据视图信息;各元数据服务器存储由视图管理模块分发的元数据视图信息,并确定的元数据服务器的网络访问地址信息;元数据服务器分层部署,最底层的元数据服务器还存储元数据内容和元数据视图信息并根据用户请求将自身存储的元数据信息提供给用户。本发明能减小带宽消耗。



CN 102833331 B

1. 一种云存储系统，其特征在于，包括元数据服务器集群、视图管理模块和客户端，其中：

客户端用于向用户提供访问所述云存储系统的接口，并根据用户的访问请求解析出用户要访问的元数据服务器的视图位置；

视图管理模块用于存储整个元数据服务器集群的全部元数据视图信息，根据所述全部元数据视图信息构建元数据服务器集群中的各元数据服务器之间的逻辑关联关系，并根据该逻辑关联关系向每个元数据服务器分发与该元数据服务器自身相关联的元数据视图信息；

元数据服务器集群中各元数据服务器用于存储由视图管理模块分发的元数据视图信息，并根据客户端或其它元数据服务器发来的元数据服务器的视图位置确定与该元数据服务器的视图位置对应的元数据服务器的网络访问地址信息；

所述元数据服务器集群的多个元数据服务器分层部署，所述元数据服务器集群中处于最底层的元数据服务器还用于存储元数据内容和视图管理模块分发的元数据视图信息，并根据用户请求将自身存储的元数据信息提供给用户。

2. 根据权利要求 1 所述的云存储系统，其特征在于，所述元数据服务器集群的每个元数据服务器均只存储了整个所述云存储系统的部分元数据内容。

3. 根据权利要求 1 所述的云存储系统，其特征在于，所述视图管理模块进一步包括数据管理单元、信息处理单元和内容视图单元，其中：

数据管理单元，用于存储所述云存储系统的全部元数据视图信息；

信息处理单元，用于根据所述数据管理单元存储的全部元数据视图信息，构建所述元数据集群的中各元数据服务器的逻辑关联关系，并对元数据视图信息进行处理，得到元数据视图信息，并向各个元数据服务器分发所述元数据视图信息，使得各个元数据服务器存储与该元数据服务器自身相关联的视图信息；

内容视图单元，用于在所述云存储系统中一元数据服务器中存储的一文件的元数据内容频繁地被另一元数据服务器访问时，将该文件写入该另一元数据服务器。

4. 根据权利要求 3 所述的云存储系统，其特征在于：

所述数据管理单元还根据不同的预设视图策略相应地存储了多套完整的与各预设视图策略分别对应的元数据视图信息；以及，

所述视图管理模块进一步包括决策单元，用于在客户端基于元数据访问请求和所述预设视图策略确定了多个用于访问的目标元数据服务器时，将所述多个用于访问目标元数据服务器之一确定为最终访问的元数据服务器。

5. 根据权利要求 4 所述的云存储系统，其特征在于，所述视图管理模块进一步包括接收单元，其中：

所述接收单元用于从各元数据服务器接收到每个最底层的元数据服务器的负载信息；

所述信息处理单元还基于元数据服务器的负载信息进行预测，得到下一时期的元数据服务器的负载预测信息；

所述决策单元，用于在客户端基于元数据访问请求和所述预设视图策略确定了多个用于访问的目标元数据服务器时，根据所述负载预测信息将所述多个用于访问目标元数据服

务器之一确定为最终访问的元数据服务器。

6. 根据权利要求 5 所述的云存储系统，其特征在于，所述决策单元进一步进行如下负载均衡控制：

当所述信息处理单元基于所述负载预测信息发现存在发生了故障或没有响应的元数据服务器发时，所述决策单元基于所述数据管理单元中存储的所述云存储系统的全部元数据视图信息来更新与发生故障的元数据服务器关联的元数据服务器的元数据视图信息，以不再向发生故障或没有响应的元数据服务器发送元数据访问请求；以及 / 或者

在由信息处理单元基于负载预测信息发现存在负载过大的元数据服务器时，所述决策单元指示该负载过大的元数据服务器停止提供元数据服务或者将全部或部分元数据复制给所述云存储系统中负载小的元数据服务器，或者所述决策单元在负载小的元数据服务器上触发生成一个所述负载过大的元数据服务器的镜像。

7. 一种云存储系统的元数据写入方法，其特征在于，

客户端在接收到用户的要写入待写元数据的写入请求时，根据预设接入策略确定用于访问接入的接入元数据服务器，所述写入请求包括待写入元数据内容；

所述客户端对所述写入请求进行预处理，以得到用于存放所述待写入元数据内容的目标元数据服务器的视图位置，其中，若所述客户端保存了与多种预设视图策略分别对应的多套元数据视图，则在所述客户端根据各种所述预设视图策略所对应的各套元数据视图信息确定了多个目标元数据服务器的视图位置时，将该多个目标元数据服务器的视图位置之一确定为最终的目标元数据服务器的视图位置；

若所确定的目标元数据服务器的视图位置对应的元数据服务器是所述接入元数据服务器，则将所述接入元数据服务器的网络访问地址信息确定为目标元数据服务器的网络访问地址信息，反之，基于所述元数据集群中各元数据服务器的层级关系，根据所述接入元数据服务器中存储的元数据视图信息和目标元数据服务器的视图位置，从所述接入元数据服务器起逐级查找目标元数据服务器，以确定目标元数据服务器的网络访问地址信息；

依据所确定的目标元数据服务器的网络访问地址信息，将该待写元数据写入所述目标元数据服务器。

8. 根据权利要求 7 所述的方法，其特征在于，所述预设接入策略为将物理地域上离所述用户最近的元数据服务器确定为所述接入元数据服务器、或者将依据所述预设视图策略对应的元数据视图信息被确定为所述用户访问最快的元数据服务器确定为所述接入元数据服务器。

9. 一种云存储系统的元数据读取的方法，其特征在于：

客户端收到用户的期望读取待元数据的读取请求时，根据预设接入策略确定用于访问接入的接入元数据服务器；

所述客户端根据所述读取请求以及所述客户端保存的元数据写入记录和元数据视图信息来确定存放了用户期望读取的元数据的目标元数据服务器的视图位置，其中，若所述客户端保存了与多种预设视图策略分别对应的多套元数据视图，则在所述客户端根据各种所述预设视图策略所对应的各套元数据视图信息确定了多个目标元数据服务器的视图位置时，将该多个目标元数据服务器的视图位置之一确定为最终的目标元数据服务器的视图位置；

若所确定的目标元数据服务器的视图位置对应的元数据服务器不是所述接入元数据服务器，则基于所述云存储系统中各元数据服务器的层级关系，根据接入元数据服务器中存储的元数据视图信息和目标元数据服务器的视图位置，从接入元数据服务器起逐级查找目标元数据服务器，以确定目标元数据服务器的网络访问地址信息，并基于目标元数据服务器的网络访问地址信息请求所述目标元数据服务器根据所述读取请求将用户所要读取的元数据内容分发到接入元数据服务器；

所述客户端从所述接入元数据服务器读取所述用户所要读取的元数据。

云存储系统及其元数据写入方法、元数据读取方法

技术领域

[0001] 本发明涉及一种云存储系统，尤其涉及用于存储大容量数据的云存储系统及其元数据内容写入方法和元数据内容读取方法。

背景技术

[0002] 元数据是“关于数据的数据”在地理空间信息中用于描述地理数据集的内容、质量、表示方式、空间参考、管理方式以及数据集的其他特征，它是实现地理空间信息共享的核心标准之一。目前，国际上对空间元数据标准内容进行研究的组织主要有三个，分别是欧洲标准化委员会 (CEN/TC287)、美国联邦地理数据委员会 (FGDC) 和国际标准化组织地理信息 / 地球信息技术委员会 (ISO/TC211)。空间元数据标准内容分两个层次。第一层是目录信息，主要用于对数据集信息进行宏观描述，它适合在数字地球的国家级空间信息交换中心或区域以及全球范围内管理和查询空间信息时使用。第二层是详细信息，用来详细或全面描述地理空间信息的空间元数据标准内容，是数据集生产者在提供空间数据集时必须要提供的信息。

[0003] 随着信息化程度的深入，信息数据量越来越大，广泛应用了元数据技术，出现了包括用于存储元数据（Metadata）的元数据服务器集群的云存储系统。

[0004] 现有的元数据服务器集群的云存储系统中，元数据服务器集群的各元数据服务器中存储整个云存储系统的元数据视图信息。然而，随着元数据内容的数据量越来越大。在非常海量的数据云存储系统中，元数据视图信息也会成为海量数据。此时，元数据服务器集群中的每个元数据服务器均需要存储所有的海量元数据视图信息，从而占用大量的存储空间。

[0005] 此外，如果元数据服务器集群中的每个元数据服务器都存储着大量的内容相同的元数据视图信息，则当客户端检索元数据内容时，在海量的元数据视图信息中搜索也将是非常耗费时间的。

发明内容

[0006] 本发明所要解决的技术问题之一是需要提供一种可减小网络的带宽消耗的云存储系统及其元数据写入方法、元数据读取方法。

[0007] 为了解决上述技术问题，本发明提供了一种云存储系统。该系统包括元数据服务器集群、视图管理模块和客户端，其中：

[0008] 客户端用于向用户提供访问所述云存储系统的接口，并根据用户的访问请求解析出用户要访问的元数据服务器的视图位置；

[0009] 视图管理模块用于存储整个元数据服务器集群的全部元数据视图信息，根据所述全部元数据视图信息构建元数据服务器集群中的各元数据服务器之间的逻辑关联关系，并根据该逻辑关联关系向每个元数据服务器分发与该元数据服务器自身相关联的元数据视图信息；

[0010] 元数据服务器集群中各元数据服务器用于存储由视图管理模块分发的元数据视图信息，并根据客户端或其它元数据服务器发来的元数据服务器的视图位置确定与该元数据服务器的视图位置对应的元数据服务器的网络访问地址信息；

[0011] 所述元数据服务器集群的多个元数据服务器分层部署，所述元数据服务器集群中处于最底层的元数据服务器还用于存储元数据内容和视图管理模块分发的元数据视图信息，并根据用户请求将自身存储的元数据信息提供给用户。

[0012] 进一步，所述元数据服务器集群的每个元数据服务器均只存储了整个所述云存储系统的部分元数据内容。

[0013] 根据本发明另一方面，所述视图管理模块进一步包括数据管理单元、信息处理单元和内容视图单元，其中：数据管理单元，用于存储所述云存储系统的全部元数据视图信息；信息处理单元，用于根据所述数据管理单元存储的全部元数据视图信息，构建所述元数据集群的中各元数据服务器的逻辑关联关系，并对元数据视图信息进行处理，得到元数据视图信息，并向各个元数据服务器分发所述元数据视图信息，使得各个元数据服务器存储与该元数据服务器自身相关联的视图信息；内容视图单元，用于在所述云存储系统中一元数据服务器中存储的一文件的元数据内容频繁地被另一元数据服务器访问时，将该文件写入该另一元数据服务器。

[0014] 根据本发明另一方面，所述数据管理单元还根据不同的预设视图策略相应地存储了多套完整的与各预设视图策略分别对应的元数据视图信息；以及，所述视图管理模块进一步包括决策单元，用于在客户端基于元数据访问请求和所述预设视图策略确定了多个用于访问的目标元数据服务器时，将所述多个用于访问目标元数据服务器之一确定为最终访问的元数据服务器。

[0015] 根据本发明另一方面，所述视图管理模块进一步包括接收单元，其中：所述接收单元用于从各元数据服务器接收到每个最底层的元数据服务器的负载信息；所述信息处理单元还基于元数据服务器的负载信息进行预测，得到下一时期的元数据服务器的负载预测信息；所述决策单元，用于在客户端基于元数据访问请求和所述预设视图策略确定了多个用于访问的目标元数据服务器时，根据所述负载预测信息将所述多个用于访问目标元数据服务器之一确定为最终访问的元数据服务器。

[0016] 根据本发明另一方面，所述决策单元进一步进行如下负载均衡控制：当所述信息处理单元基于所述负载预测信息发现存在发生了故障或没有响应的元数据服务器发时，所述决策单元基于所述数据管理单元中存储的所述云存储系统的全部元数据视图信息来更新与发生故障的元数据服务器关联的元数据服务器的元数据视图信息，以不再向发生故障或没有响应的元数据服务器发送元数据访问请求；以及／或者在由信息处理单元基于负载预测信息发现存在负载过大的元数据服务器时，所述决策单元指示该负载过大的元数据服务器停止提供元数据服务或者将全部或部分元数据复制给所述云存储系统中负载小的元数据服务器，或者所述决策单元在负载小的元数据服务器上触发生成一个所述负载过大的元数据服务器的镜像。

[0017] 根据本发明又一方面，还提供一种云存储系统的元数据写入方法。该方法包括：

[0018] 客户端在接收到用户的要写入待写元数据的写入请求时，根据预设接入策略确定用于访问接入的接入元数据服务器，所述写入请求包括待写入元数据内容；

[0019] 所述客户端对所述写入请求进行预处理，以得到用于存放所述待写入元数据内容的目标元数据服务器的视图位置；

[0020] 若所确定的目标元数据服务器的视图位置对应的元数据服务器是所述接入元数据服务器，则将所述接入元数据服务器的网络访问地址信息确定为目标元数据服务器的网络访问地址信息，反之，基于所述元数据集群中各元数据服务器的层级关系，根据所述接入元数据服务器中存储的元数据视图信息和目标元数据服务器的视图位置，从所述接入元数据服务器起逐级查找目标元数据服务器，以确定目标元数据服务器的网络访问地址信息；

[0021] 依据所确定的目标元数据服务器的网络访问地址信息，将该待写元数据写入所述目标元数据服务器。

[0022] 进一步，所述预设接入策略为将物理地域上离所述用户最近的元数据服务器确定为所述接入元数据服务器、或者将依据所述预设视图策略对应的元数据视图信息被确定为所述用户访问最快的元数据服务器确定为所述接入元数据服务器。

[0023] 进一步，若所述客户端保存了与多种所述预设视图策略分别对应的多套元数据视图，则在所述客户端根据各种所述预设视图策略所对应的各套元数据视图信息确定了多个目标元数据服务器的视图位置时，将该多个目标元数据服务器的视图位置之一确定为最终的目标元数据服务器的视图位置；依据所确定的最终的目标元数据服务器的网络访问地址信息，将该待写元数据写入所述目标元数据服务器。

[0024] 根据本发明的又一方面，还提供了一种云存储系统的元数据读取的方法。该方法包括：客户端收到用户的期望读取待元数据的读取请求时，根据预设接入策略确定用于访问接入的接入元数据服务器；所述客户端根据所述读取请求以及所述客户端保存的元数据写入记录和元数据视图信息来确定存放了用户期望读取的元数据的目标元数据服务器的视图位置；若所确定的目标元数据服务器的视图位置对应的元数据服务器不是所述接入元数据服务器，则基于所述云存储系统中各元数据服务器的层级关系，根据接入元数据服务器中存储的元数据视图信息和目标元数据服务器的视图位置，从接入元数据服务器起逐级查找目标元数据服务器，以确定目标元数据服务器的网络访问地址信息，并基于目标元数据服务器的网络访问地址信息请求所述目标元数据服务器根据所述读取请求将用户所要读取的元数据内容分发到接入元数据服务器；所述客户端从所述接入元数据服务器读取所述用户所要读取的元数据。

[0025] 与现有技术相比，本发明的一个或多个实施例可以具有如下优点：

[0026] 根据本发明实施例提供的云存储系统能较好地应用于视图信息量大的云存储系统，由分层设置的各元数据服务器根据各自所存储的元数据视图来相互协调地实现元数据访问，相比传统技术，可减小网络的带宽消耗，满足网络流量和元数据服务器的视图均衡的要求。

[0027] 本发明的其他优点、目标，和特征在某种程度上将在随后的说明书中进行阐述，并且在某种程度上，基于对下文的考察研究对本领域技术人员而言将是显而易见的，或者可以从本发明的实践中得到教导。本发明的目标和其他优点可以通过下面的说明书，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

[0028] 附图用来提供对本发明的进一步理解，并且构成说明书的一部分，与本发明的实施例共同用于解释本发明，并不构成对本发明的限制。在附图中：

[0029] 图 1 是根据本发明实施例一的云存储系统的结构示意图；

[0030] 图 2 是根据本发明实施例二的云存储系统的元数据写入方法的流程图；

[0031] 图 3 是根据本发明实施例三的云存储系统的元数据读取方法的流程图。

具体实施方式

[0032] 为使本发明的目的、技术特征和实施效果更加清楚，下面将结合附图及具体实施例对本发明的实施例进行详细描述。需要说明的是，只要不构成冲突，本发明中的各个实施例以及各实施例中的各个特征可以相互结合，所形成的技术方案均在本发明的保护范围之内。

[0033] 另外，在附图的流程图示出的步骤可以在诸如一组计算机可执行指令的计算机系统中执行，并且，虽然在流程图中示出了逻辑顺序，但是在某些情况下，可以以不同于此处的顺序执行所示出或描述的步骤。

[0034] 以下本发明实施例中提到的内容即为用户需要访问的元数据内容，本发明实施例中提到的文件，即为包含元数据内容的文件。

[0035] 实施例一

[0036] 图 1 描述了本发明实施例提供的一种云存储系统的结构示意图。该系统包括：元数据服务器集群 10、视图管理模块 20 和客户端 30。

[0037] 元数据服务器集群 10 的多个元数据服务器分层部署。元数据服务器集群 10 的服务器可以通过网络来实现合作，即相互请求服务和提供服务。

[0038] 元数据服务器集群 10 中各元数据服务器用于存储由视图管理模块 20 分发的视图信息，并根据客户端 30 或其它元数据服务器发来的请求、查询指定的元数据服务器的网络访问地址信息。该网络访问地址信息可为例如网址、域名等使得客户端 30 能够访问元数据服务器的网络地址信息。

[0039] 进一步，元数据服务器集群 10 中处于最底层的元数据服务器还用于存储元数据内容(亦称作元数据信息或元数据)和视图管理模块 20 分发的视图信息，并根据用户请求将自身存储的元数据信息提供给用户。需要说明的是，每个元数据服务器均没有存储整个云存储系统的全部元数据内容(即，每个元数据服务器只存储了整个所述云存储系统的部分元数据内容)，各个元数据服务器存储的元数据内容的并集是全部元数据内容。

[0040] 视图管理模块 20，用于存储整个元数据服务器集群 10 的全部元数据视图信息（亦将全部元数据视图信息称为视图索引信息），以及根据该视图索引信息，构建元数据服务器集群 10 中的各元数据服务器之间的逻辑关联关系（或称为逻辑连接关系），并根据该逻辑关联关系向每个元数据服务器分发与该元数据服务器自身相关联的视图信息。与某个元数据服务器自身相关联的视图信息可以包括该元数据服务器自身的名称、地理空间信息及网络访问地址信息、与该元数据服务器逻辑关联的其它元数据服务器的名称、地理空间信息及网络访问地址信息等。

[0041] 优选地，视图管理模块 20 还可周期性地检测元数据服务器集群 10 中每个元数据服务器的负载信息，可优选地只周期性检测每个最底层元数据服务器的负载信息。这样，便

于云存储系统将负载压力过大的元数据服务器的服务切换到相对负载小的其他元数据服务器上。所述周期可以设为 10 分钟、一小时、一天等，可以根据元数据内容的访问频率的变化快慢来确定，如访问频率变化较快则周期可以设短一些，访问频率变化较慢则周期可以设长一些。

[0042] 优选地，视图管理模块 20 还可根据不同的预设视图策略，构建元数据服务器集群 10 中的各元数据服务器之间的多套逻辑关联关系，并根据该多套逻辑关联关系向每个元数据服务器分发与该元数据服务器自身相关联的多套视图信息。各套视图信息与各套逻辑关联关系相对应。

[0043] 这样，各元数据服务器存储着与各套逻辑关联关系对应的各套与该元数据服务器自身相关联的视图信息。当客户端 30 要通过访问存储着多套视图信息的元数据服务器来确定要访问的元数据服务器的网络访问地址信息时，要先基于各套与该元数据服务器自身相关联的视图信息分别确定一网络访问地址信息，然后再由决策单元最终确定所要用到的网络访问地址信息，从而尽量将元数据内容分发给云存储系统中最合适的元数据服务器。预设视图策略可以多种多样，例如，预设视图策略可以根据实际地域分布，如中国下面分华北、华东、华南…，华北下面分北京、天津、河北…，北京下面分海淀区、朝阳区、东城区、西城区。预设视图策略还可以根据文件种类，如分为工作、娱乐、学习等，工作下面分技术、经济、人文等，娱乐下面分音乐、体育、电影等。还可以根据业务的不同，拆成不同的逻辑关系。

[0044] 如图 1 所述，第 n 层元数据服务器在接收到来自客户端 30 的元数据访问请求时，根据自身存储的元数据视图信息来获取客户端 30 所要访问的元数据服务器的网络访问地址信息（详见实施例二）。

[0045] 在本发明中，第 n 层元数据服务器作为元数据内容的实际存储位置；每个第 n 层元数据服务器没有存储全部元数据信息，所有第 n 层元数据服务器存储的内容之和是完整的元数据内容。

[0046] 需要说明的是，视图管理模块 20 可以设置在某个元数据服务器上，也可以是单独的逻辑模块。

[0047] 此外，所述元数据视图信息既可以是元数据服务器的地域信息（地理空间信息），也可以是普通的树形结构信息，或者用户自定义的逻辑关系信息，或根据特定的方法（比如 hash 算法）构造的元数据视图信息。

[0048] 客户端 30 用于向用户提供访问所述云存储系统的接口，解析用户的读写访问请求，该访问请求包含要访问的元数据服务器的视图位置信息（简称视图位置）、文件标识、文件大小等信息。此外，客户端 30 还保存关于写入元数据的写入信息记录。视图位置信息可以为需要访问的元数据在元存储网络中的网络访问地址，如果按地域构建的视图信息，可以为中国 - 华北 - 北京 - 海淀具体元数据文件。

[0049] 较佳的，所述视图管理模块 20 可包括：数据管理单元、信息处理单元、接收单元、决策单元和内容视图单元。

[0050] 数据管理单元，用于存储云存储系统的全部元数据视图信息。优选地，可以根据不同的预设视图策略相应地存储多套完整的与各预设视图策略分别对应的元数据视图信息，这样，使得各元数据服务器存储与不同的预设视图策略一一对应的多套元数据视图信息。在这种情况下，各元数据服务器中存储的各套元数据视图信息均不是整套的元数据视图信

息，而是各个整套（全部）元数据视图信息中与自己相关的那部分元数据视图信息。

[0051] 信息处理单元，用于根据数据管理单元存储的全部元数据视图信息，构建元数据集群中各元数据服务器的逻辑关联关系，并对元数据视图信息进行处理，得到要向各个元数据服务器分发的元数据视图信息，并进行分发，使得各个元数据服务器存储与该元数据服务器自身相关联的视图信息。

[0052] 优选地，该云存储系统还可包括接收单元。接收单元从各元数据服务器接收到每个最底层的元数据服务器的负载信息。进一步，可由信息处理单元基于元数据服务器的负载信息进行预测，得到下一时期的元数据服务器的负载预测信息。例如，根据预先设定的负载均衡策略，按预先设置的权重整合多个历史时期的状态数据得到对下一时期的元数据服务器的负载预测信息。

[0053] 所述负载信息可以包括：云存储系统中各个元数据服务器上的各个元数据的访问频率，以及各个元数据服务器的可用存储容量、出口带宽、入口带宽和服务视图容量（可用存储空间）、特定的负载算法和响应时延等。

[0054] 决策单元，用于在客户端 30 基于用户的元数据访问请求和前述预设视图策略确定了多个用于访问的目标元数据服务器时，将多个用于访问目标元数据服务器之一确定为最终访问的元数据服务器。其中，用户的元数据访问请求可以为读取请求或写入请求等。这样，决策单元可综合几种预设视图策略来实现元数据更合理分发。

[0055] 例如，决策单元可根据信息处理单元处理得到的负载预测信息，将上述多个目标元数据服务器之一确定为要访问的元数据服务器并将其通知客户端 30。也就是说，当客户端 30 根据多种预设视图策略确定了不同的目标元数据服务器时，客户端 30 不知道最终要访问哪个元数据服务器，因此与决策单元进行交互，由决策单元来通知客户端 30 最终要访问这些目标元数据服务器中的哪一个。

[0056] 此外，当信息处理单元基于负载预测信息发现存在发生了故障或没有响应的元数据服务器发时，决策单元进行负载均衡控制。更具体地，可由决策单元动态地基于所述数据管理单元中存储的所述云存储系统的全部元数据视图信息来更新与该发生故障的元数据服务器关联的元数据服务器的元数据视图信息，以不再向那个发生了故障或没有响应的元数据服务器发送元数据访问请求，例如，可以将元数据视图信息中关于发生故障的元数据服务器的信息删除。

[0057] 此外，在由信息处理单元基于负载预测信息发现存在负载过大的元数据服务器时，可以由决策单元进行负载均衡控制，例如，决策单元可自动指示该负载过大的元数据服务器停止提供元数据服务或者将其全部或部分元数据复制给该系统中负载较小的元数据服务器，或者可通过决策单元在负载较小的元数据服务器上触发生成一个该负载过大的元数据服务器的镜像，从而由该负载较小的元数据服务器替代该负载过大的元数据服务器来供服务。

[0058] 内容视图单元，用于根据所述元数据内容的视图位置信息，将元数据内容分发给所述云存储系统中合适视图的元数据服务器。

[0059] 更具体地，内容视图单元可以在云存储系统中某一元数据服务器中存储的一文件的元数据内容频繁地被另一元数据服务器访问时，将该文件写入另一元数据服务器，并将该文件写入所述的另一元数据服务器的路径信息反馈给客户端 30。这样，系统可更快的对

客户提供元数据文件访问服务。

[0060] 进一步，为了保证数据的一致性，系统可对元数据服务器进行分组，使得组内的元数据服务器之间进行数据备份。如有500个元数据服务器，5个为一组，组内进行数据备份，保证数据的一致性，这样不但保证了数据的可靠性，同时还提升了对外访问的效率、性能。

[0061] 本发明实施例提供了一种云存储系统，应用于视图信息量大的云存储系统，由分层设置的各元数据服务器根据自各所存储的元数据视图来相互协调地实现元数据访问，相比传统技术，可减小网络的带宽消耗，提高用户访问元数据内容的命中率，满足网络流量和元数据服务器的视图均衡的要求。

[0062] 实施例二

[0063] 前述实施例中关于云存储系统的相关说明，同样适应于本实施例。图2是本发明实施例提供的一种云存储系统的元数据写入方法的流程图。下面结合图2来详细说明本发明实施例二的云存储系统的元数据写入方法。

[0064] 步骤S210，客户端30在接收到用户的要写入待写元数据的写入请求时，根据预设接入策略确定用于访问接入的元数据服务器(简称接入元数据服务器)。

[0065] 该预设接入策略可以为例如将物理地域上离该用户最近的元数据服务器确定为接入元数据服务器、或将按前述预设视图策略对应的视图信息确定该用户访问最快的元数据服务器确定为接入元数据服务器等。该接入元数据服务器优选为处于最底层的元数据服务器。

[0066] 该写入请求包括要写入的元数据内容(简称为待写入元数据内容)。

[0067] 步骤S220，客户端30可对写入请求进行预处理，以用于存放该待写入元数据内容的目标元数据服务器的视图位置。

[0068] 可选地，步骤S220还可进一步包括：若该客户端30保存了与多种预设视图策略分别对应的多套元数据视图，则在客户端30根据各种预设视图策略所对应的各套元数据视图信息确定了多个目标元数据服务器的视图位置时，客户端30可利用决策单元来将该多个目标元数据服务器的视图位置之一确定为最终的目标元数据服务器的视图位置，从而使元数据访问量可更均衡地分布到各个元数据服务器。

[0069] 步骤S230，判断在步骤S220中确定的目标元数据服务器的视图位置所对应的元数据服务器(出现多个目标元数据服务器的视图位置时，该目标元数据服务器的视图位置指最终的接入元数据服务器的视图位置)是否为步骤S210中确定的接入元数据服务器。如果是，则进入步骤S240，反之，进入步骤S250。

[0070] 步骤S240，将接入元数据服务器的网络访问地址信息确定为目标元数据服务器的网络访问地址信息。

[0071] 步骤S250基于各元数据服务器的层级关系，根据接入元数据服务器中存储的视图信息和目标元数据服务器的视图位置，从接入元数据服务器起逐级查找目标元数据服务器，以确定目标元数据服务器的网络访问地址信息。

[0072] 更具体地，根据接入元数据服务器中存储的视图信息确定与该接入元数据服务器中存储的视图信息关联的元数据服务器(简称一级关联服务器)，若一级关联服务器中存在该目标元数据服务器，则根据接入元数据服务器中存储的视图信息来确定目标元数据服务器的网络访问地址信息；反之，进一步根据(多个)一级关联服务器中存储的视图信息，来确

定与一级关联服务器相关联的元数据服务器(简称二级关联服务器),若二级关联服务器中存在该目标元数据服务器,则根据一级关联服务器中存储的视图信息来确定目标元数据服务器的网络访问地址信息,依此类推,在视图信息中查找到该目标元数据服务器,从而确定该目标元数据服务器的网络访问地址信息。

[0073] 步骤 S260,依据步骤 S240 或 S250 中确定的该目标元数据服务器的网络访问地址信息,将该待写元数据写入该目标元数据服务器。

[0074] 可选地,步骤 S270,目标元数据服务器可将结果返回给客户端 30,客户端 30 可保存与写入元数据相关的元数据写入记录。这样,提高用户在从云存储系统中读取元数据时确定目标元数据服务器的视图位置和 / 或接入元数据服务器的效率。

[0075] 综上所述,通过数据服务器集群中分层设置的元数据服务器来协助查找要最后要写入的目标元数据服务器的网络访问地址信息,可以将确定目标元数据服务器的网络访问地址信息的处理分布到各级的元数据服务器来进行,从而使负载更均衡。

[0076] 实施例三

[0077] 前述实施例中关于云存储系统的相关说明,同样适应于本实施例。图 3 是本发明实施例提供的一种云存储系统的元数据读取的方法的流程图。下面结合图 3 来详细说明本发明实施例三的云存储系统的元数据读取方法。

[0078] 步骤 S310,客户端 30 收到用户的期望读取待元数据的读取请求时,根据预设接入策略确定用于访问接入的元数据服务器(简称接入元数据服务器)。该步骤中确定接入元数据服务器的方式与步骤 S210 中类似,在此不再详细展开说明。

[0079] 该读取请求可包括用户期望读取的元数据的属性。用户期望读取的元数据的属性可以为用户期望读取的元数据的一个或多个字段的值。

[0080] 步骤 S320,客户端 30 根据该读取请求以及该客户端 30 保存的元数据写入记录和视图信息来确定存放了用户期望读取的元数据的目标元数据服务器的视图位置。更具体地,基于所保存的元数据写入记录及视图信息来对待读取元数据内容进行预处理,得到存放了用户期望读取的元数据的目标元数据服务器的视图位置。

[0081] 步骤 S330,判断在步骤 S320 中确定的目标元数据服务器的视图位置所对应的元数据服务器是否为步骤 S310 中确定的接入元数据服务器。如果是,则进入步骤 S360,反之,进入步骤 S340。

[0082] 步骤 S340,基于各元数据服务器的层级关系,根据接入元数据服务器中存储的视图信息和目标元数据服务器的视图位置,从接入元数据服务器起逐级查找目标元数据服务器,以确定目标元数据服务器的网络访问地址信息,然后进入步骤 S350。该步骤的处理与步骤 S250 类似,在此不再展开说明。

[0083] 步骤 S350,基于目标元数据服务器的网络访问地址信息请求目标元数据服务器根据该读取请求将用户所要读取的元数据内容分发到接入元数据服务器。

[0084] 步骤 S360,客户端 30 从接入元数据服务器读取用户所要读取的元数据。

[0085] 综上所述,通过在接入元数据服务器不是目标元数据服务器时、从接入元数据服务器起逐级查找目标元数据服务器以确定目标元数据服务器的网络访问地址信息,这样较好的减少单个元数据服务器存储的视图信息的数据量,从而提高搜索效率。

[0086] 本领域的技术人员应该明白,上述的本发明的各模块或各步骤可以用通用的计算

装置来实现，它们可以集中在单个的计算装置上，或者分布在多个计算装置所组成的网络上，可选地，它们可以用计算装置可执行的程序代码来实现，从而，可以将它们存储在存储装置中由计算装置来执行，或者将它们分别制作成各个集成电路模块，或者将它们中的多个模块或步骤制作成单个集成电路模块来实现。这样，本发明不限制于任何特定的硬件和软件结合。

[0087] 虽然本发明所揭露的实施方式如上，但所述的内容只是为了便于理解本发明而采用的实施方式，并非用以限定本发明。任何本发明所属技术领域内的技术人员，在不脱离本发明所揭露的精神和范围的前提下，可以在实施的形式上及细节上作任何的修改与变化，但本发明的专利保护范围，仍须以所附的权利要求书所界定的范围为准。

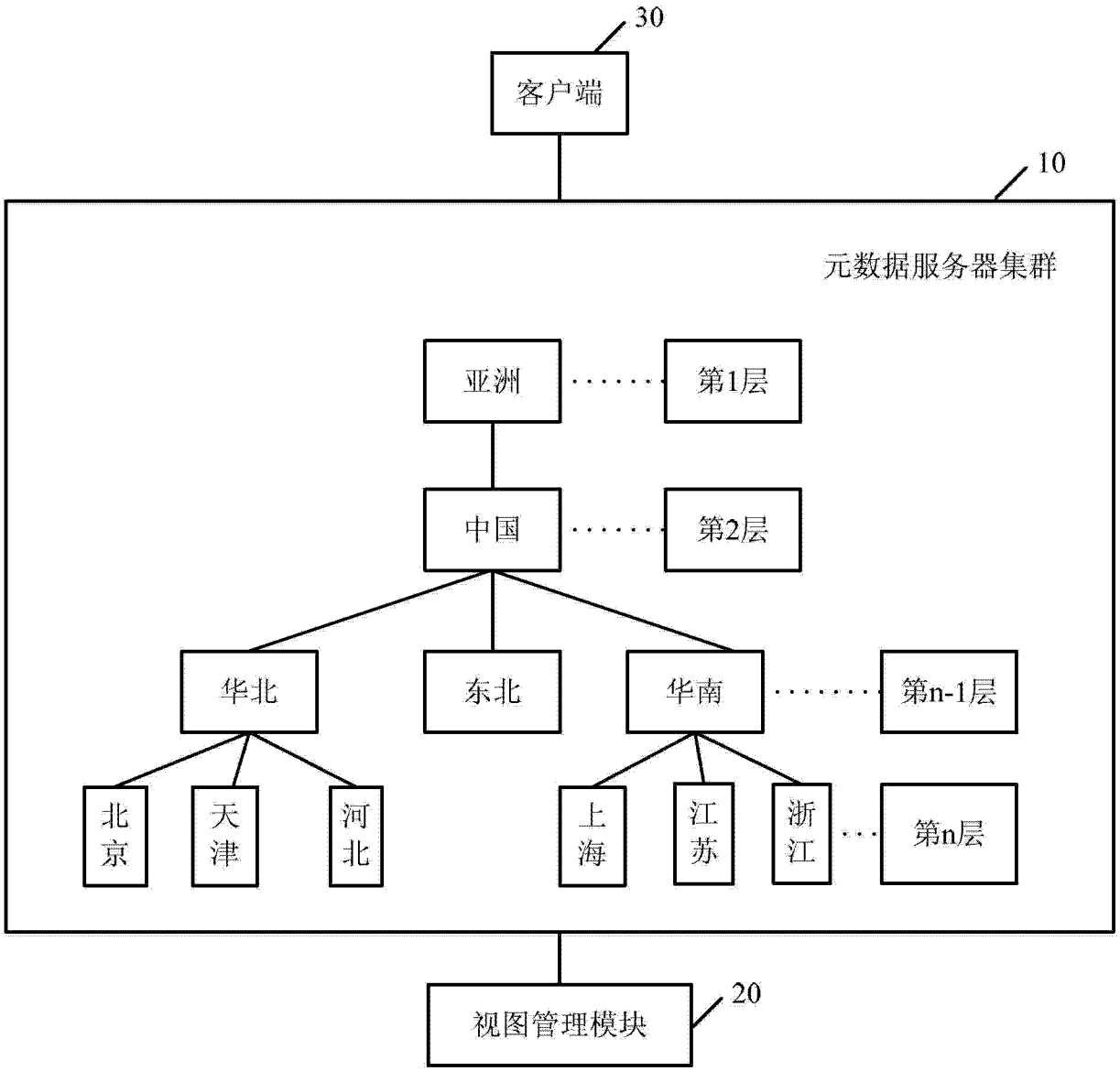

30
客户端
10
元数据服务器集群
亚洲
第1层
中国
第2层
华北
东北
华南
第n-1层
北京
天津
河北
上海
江苏
浙江
第n层
视图管理模块
20

图 1

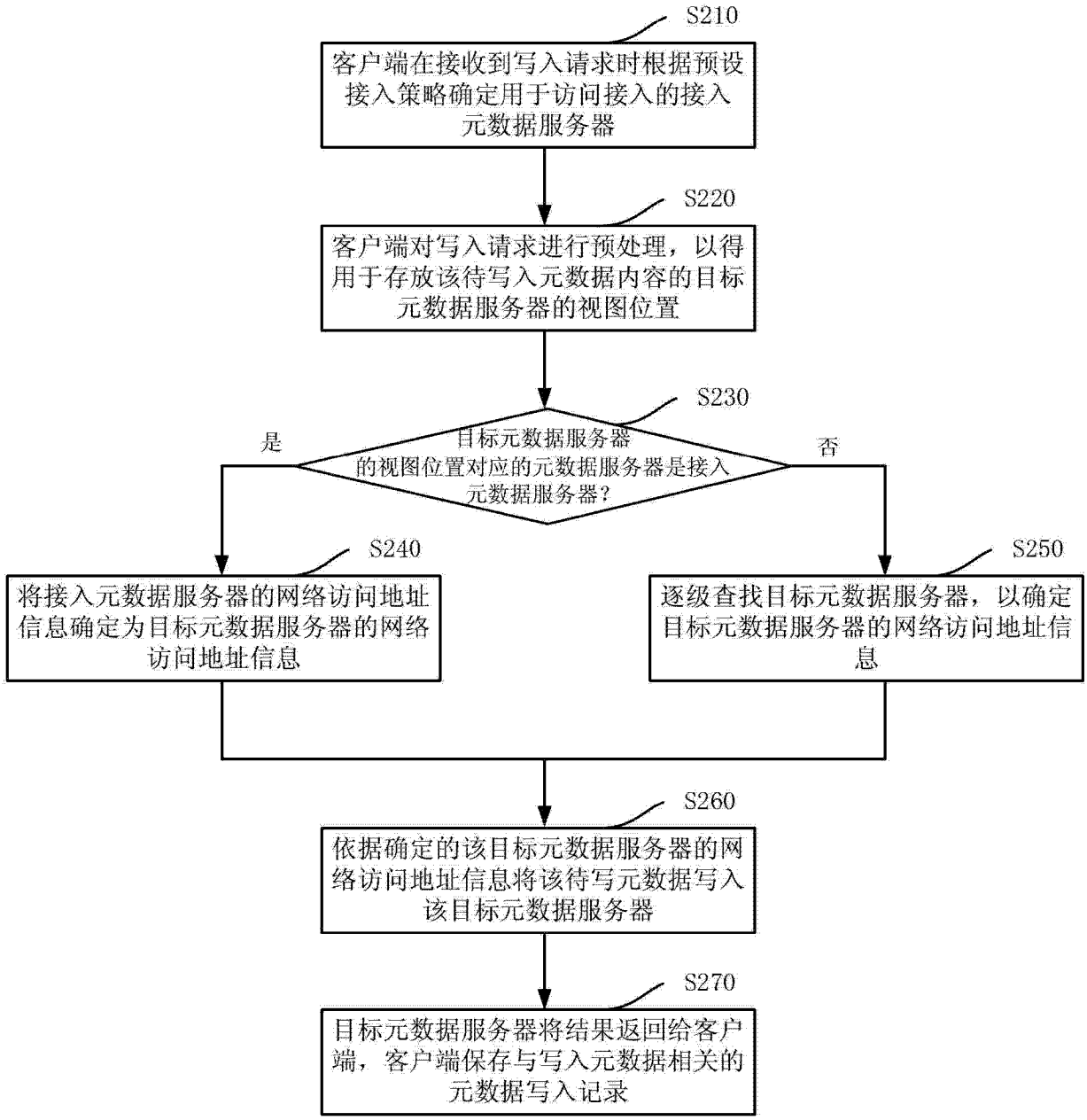

S210
客户端在接收到写入请求时根据预设接入策略确定用于访问接入的接入元数据服务器
S220
客户端对写入请求进行预处理，以得用于存放该待写入元数据内容的目标元数据服务器的视图位置
S230
目标元数据服务器的视图位置对应的元数据服务器是接入元数据服务器？
是
否
S240
将接入元数据服务器的网络访问地址信息确定为目标元数据服务器的网络访问地址信息
S250
逐级查找目标元数据服务器，以确定目标元数据服务器的网络访问地址信息
S260
依据确定的该目标元数据服务器的网络访问地址信息将该待写元数据写入该目标元数据服务器
S270
目标元数据服务器将结果返回给客户端，客户端保存与写入元数据相关的元数据写入记录

图 2

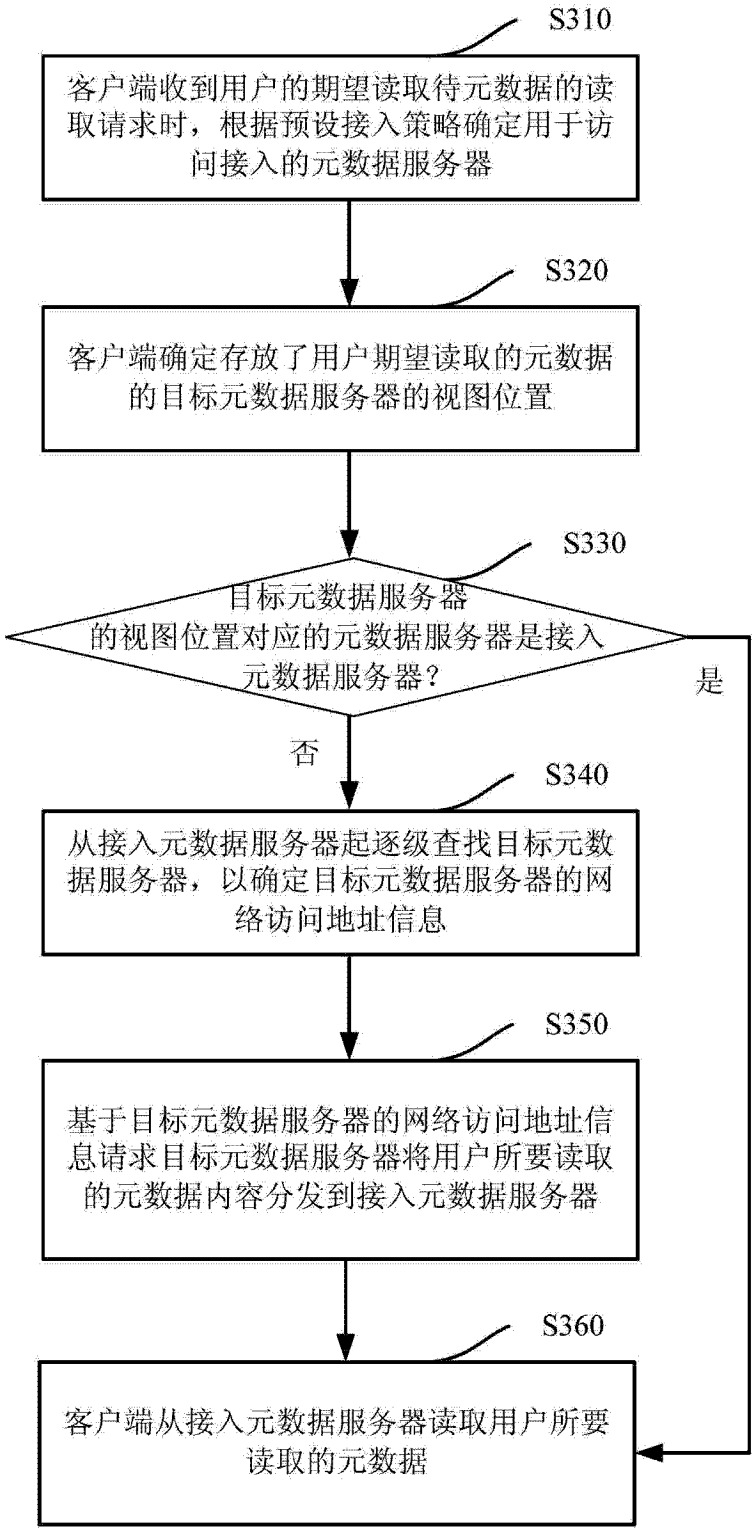

S310
客户端收到用户的期望读取待元数据的读取请求时，根据预设接入策略确定用于访问接入的元数据服务器
S320
客户端确定存放了用户期望读取的元数据的目标元数据服务器的视图位置
S330
目标元数据服务器的视图位置对应的元数据服务器是接入元数据服务器？
是
否
S340
从接入元数据服务器起逐级查找目标元数据服务器，以确定目标元数据服务器的网络访问地址信息
S350
基于目标元数据服务器的网络访问地址信息请求目标元数据服务器将用户所要读取的元数据内容分发到接入元数据服务器
S360
客户端从接入元数据服务器读取用户所要读取的元数据

图 3