(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 103279568 A
(43)申请公布日 2013.09.04

(21)申请号 201310242787.8

(22)申请日 2013.06.18

(71)申请人 无锡紫光存储系统有限公司
地址 214000 江苏省无锡市新区净慧东道 77 号 -8-3-4

(72)发明人 周海波 苗东 周泉 于强

(74)专利代理机构 北京品源专利代理有限公司 11332
代理人 马晓亚

(51)Int.Cl.
G06F 17/30(2006.01)
G06F 11/30(2006.01)

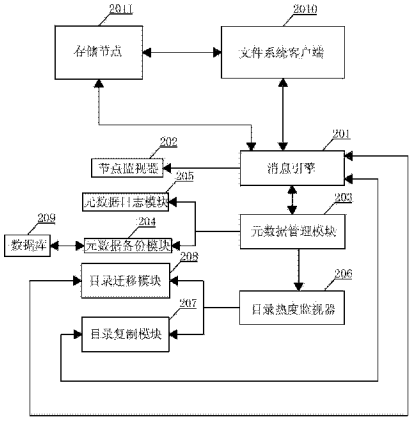
权利要求书3页 说明书6页 附图3页

(54)发明名称

一种元数据管理系统及方法

(57)摘要

本发明公开一种元数据管理系统和方法，该系统包括：消息引擎、节点监视器、元数据管理模块、元数据备份模块、元数据日志模块、目录热度监视器、目录复制模块以及目录迁移模块。本发明中元数据分布在元数据集群中的节点中，而不是内存容量受限的单一容量中，使得存储容量可横向扩展，并且元数据的分布使用动态分布策略，提高了分布式文件系统在企业应用环境所需的高可用性、存储容量以及访问性能的高可扩展性。



CN 103279568 A

1. 一种元数据管理系统，其特征在于，包括：消息引擎、节点监视器、元数据管理模块、元数据备份模块、元数据日志模块、目录热度监视器、目录复制模块以及目录迁移模块；

所述消息引擎用于负责与文件系统客户端和存储节点之间通信；

所述节点监视器用于维护元数据集群中的每个节点的健康信息；

所述元数据管理模块用于根据文件系统客户端的文件访问类型，完成对目录以及文件对像的操作请求；

所述元数据备份模块用于根据元数据的备份要求，将相关的目录对象对应的元数据备份到元数据集群中的其它元数据节点中；

所述元数据日志模块用于将文件系统客户端对文件系统操作的日志保存到本地；

所述目录热度监视器用于实时监测目录对象访问的频繁程度，并根据其它元数据节点的负载情况，对目录对象是否需要复制和迁移进行决策；

所述目录复制模块用于根据目录热度监视器发布的目录对象复制要求，完成目录对象向其它元数据节点的复制；

所述目录迁移模块用于根据目录热度监视器发布的目录对象迁移要求，完成目录对象向其它元数据节点的迁移。

2. 根据权利要求 1 所述的元数据管理系统，其特征在于，所述消息引擎具体用于：

负责与文件系统客户端和存储节点之间通信，当文件系统客户端进行文件操作时，返回元数据信息，或者数据块映射信息，或者操作结果，并接收来自于存储节点定期上传的数据块映射信息。

3. 根据权利要求 1 所述的元数据管理系统，其特征在于，所述每个节点的健康信息包括元数据节点的工作负载、访问热度以及心跳情况。

4. 根据权利要求 1 至 3 之一所述的元数据管理系统，其特征在于，所述对目录以及文件对像的操作请求包括：新建目录、删除目录、读取目录、创建文件、删除文件以及重命名文件名称或者目录名称。

5. 一种元数据管理方法，其特征在于，包括如下步骤：

A、根据元数据节点的负载，计算目录子树的访问频度，并根据该访问频度择出需要迁移的目录子树；

B、将需要迁移的目录子树迁移到负载较低的元数据节点上，达到元数据集群负载均衡；

C、元数据节点根据目录的访问热度，将目录访问热度符合复制条件的目录子树复制到其它的元数据节点中形成副本。

6. 根据权利要求 5 所述的元数据管理方法，其特征在于，所述步骤 A 具体包括：

A1、元数据节点通过节点监视器获取其它所有元数据节点的负载信息，并由负载最高的元数据节点计算出当前元数据集群的负载均衡度；当负载均衡度达到规定阈值时，执行步骤 A2；

A2、通过节点监视器获取其它所有元数据节点的负载信息和目录访问热度信息，目录热度监视器计算目录需要迁移的阈值以及目录子树迁移的目标元数据节点集合。

7. 根据权利要求 6 所述的元数据管理方法，其特征在于，所述步骤 B 具体包括：

B1、在目录子树迁移的目标元数据节点集合中，选取目录访问热度最低的元数据节点，

查询该元数据节点的缓存中是否存在从目录访问热度最低的元数据节点迁入的目录，如果符合目录子树迁移条件，则将该类目录子树迁移到目录访问热度最低的元数据节点中；否则，在该元数据节点的目录中选择相应热度的目录子树迁移到目录访问热度最低的元数据节点中；

B2、如果没有目录子树达到迁移的条件，则从目标子树迁移的元数据节点集合中选择下一个元数据节点，重复执行步骤B1。

8. 根据权利要求7所述的元数据管理方法，其特征在于，所述步骤B1具体包括：

B101、源元数据节点的目录迁移模块接收到目录子树迁移的消息后，冻结该目录子树，并向目标元数据节点发送需要迁移目录子树的路径消息；

B102、目标元数据节点获取需要迁移的目录子树到根目录路径上所有的目录节点的相关信息，创建缓存，并向源元数据节点发送确认信息；

B103、源元数据节点将要迁移的目录子树中的根节点以及其所有叶子结点中的目录节点数据打包发送到目标元数据节点；

B104、目标元数据节点将解包后的目录子树副本数据保存到步骤B102创建的缓存中；

B105、如果该被迁移的目录子树不存在副本，则跳转到步骤B107；否则，向所有拥有被迁移的目录子树副本的元数据节点发送目录子树迁移通知消息；

B106、所述拥有被迁移的目录子树副本的元数据节点修改被迁移目录子树的迁移状态，并发送确认消息；

B107、源元数据节点修改被迁移的目录子树的状态，并将目录子树中所有节点的导出信息打包发送到目标元数据节点中；

B108、目标元数据节点解包目录子树的导出信息，将其缓存到本地，并设置元数据信息的主副状态，再在副本集合中增加源元数据节点的记录，最后向源元数据节点发送确认信息；

B109、在源元数据节点上，如果该被迁移的目录子树不存在副本，则跳转到步骤B1011；否则，向所有拥有被迁移的目录子树副本的元数据节点发送导出通知信息，并修改迁移目录的迁移状态；

B1010、拥有被迁移的目录子树副本的元数据节点修改被迁移的目录子树的迁移状态，并向源元数据节点发送确认消息；

B1011、源元数据节点向目标元数据节点发送目录子树导出的完成消息，并解冻被迁移的目录子树；

B1012、目标元数据节点解冻被迁移的目录子树，迁移过程结束。

9. 根据权利要求8所述的元数据管理方法，其特征在于，所述步骤A2具体包括：

A201、源元数据节点根据元数据节点的负载信息，选择n个负载大小满足预设要求并且元数据服务处于正常状态的元数据节点作为目录子树的备份元数据节点，并向这些被备份节点发送目录子树复制的通知信息，然后将该目录子树的访问热度调整为原有热度的1/n；

A202、备份元数据节点向源元数据节点发送复制目录子树的查询消息；

A203、源元数据节点将被查询的目录子树的相关数据打包发送给备份元数据节点；

A204、备份元数据节点将目录子树解包，将目录子树的相关数据加载到本地缓存，修改

复制目录的副本信息。

一种元数据管理系统及方法

技术领域

[0001] 本发明涉及分布式文件系统中元数据管理技术领域，尤其涉及一种元数据管理系统和方法。

背景技术

[0002] 在目前的支持大数据的分布式文件系统中，采用的是单节点的元数据服务模型。首先，如果由于硬件故障或者软件错误导致元数据节点失效，进行文件操作时将无法获得元数据信息，文件系统将无法继续使用，导致整个系统瘫痪。尽管有些技术方案使用了Active-Backup（备份算法)模式增加了备份节点，但是当主节点发生故障后，元数据服务切换到备份节点时的这个过程需要较长的时间，不能满足企业应用要求的高可用性。其次，由于元数据保存在元数据节点的内存中，而元数据节点的内存是有限的，在大数据的应用中，文件系统的规模不断增大，元数据的大小急速增加，元数据节点的内存将无法容纳如此多的元数据，从而限制了文件系统的存储容量；另外，当访问文件系统中的数据时，所有的访问都需要向元数据节点发送请求。当文件系统的访问量增大时，元数据节点就成为文件访问的性能瓶颈。

发明内容

[0003] 本发明的目的在于通过一种元数据管理系统和方法，来解决以上背景技术部分提到的问题。

[0004] 为达此目的，本发明采用以下技术方案：

[0005] 一种元数据管理系统，其包括：消息引擎、节点监视器、元数据管理模块、元数据备份模块、元数据日志模块、目录热度监视器、目录复制模块以及目录迁移模块；

[0006] 所述消息引擎用于负责与文件系统客户端和存储节点之间通信；

[0007] 所述节点监视器用于维护元数据集群中的每个节点的健康信息；

[0008] 所述元数据管理模块用于根据文件系统客户端的文件访问类型，完成对目录以及文件对像的操作请求；

[0009] 所述元数据备份模块用于根据元数据的备份要求，将相关的目录对象对应的元数据备份到元数据集群中的其它元数据节点中；

[0010] 所述元数据日志模块用于将文件系统客户端对文件系统操作的日志保存到本地；

[0011] 所述目录热度监视器用于实时监测目录对象访问的频繁程度，并根据其它元数据节点的负载情况，对目录对象是否需要复制和迁移进行决策；

[0012] 所述目录复制模块用于根据目录热度监视器发布的目录对象复制要求，完成目录对象向其它元数据节点的复制；

[0013] 所述目录迁移模块用于根据目录热度监视器发布的目录对象迁移要求，完成目录对象向其它元数据节点的迁移。

[0014] 特别地，所述消息引擎具体用于：

[0015] 负责与文件系统客户端和存储节点之间通信，当文件系统客户端进行文件操作时，返回元数据信息，或者数据块映射信息，或者操作结果，并接收来自于存储节点定期上传的数据块映射信息。

[0016] 特别地，所述每个节点的健康信息包括元数据节点的工作负载、访问热度以及心跳情况。

[0017] 特别地，所述对目录以及文件对像的操作请求包括：新建目录、删除目录、读取目录、创建文件、删除文件以及重命名文件名称或者目录名称。

[0018] 本发明还公开了一种元数据管理方法，其包括如下步骤：

[0019] A、根据元数据节点的负载，计算目录子树的访问频度，并根据该访问频度择出需要迁移的目录子树；

[0020] B、将需要迁移的目录子树迁移到负载较低的元数据节点上，达到元数据集群负载均衡；

[0021] C、元数据节点根据目录的访问热度，将目录访问热度符合复制条件的目录子树复制到其它的元数据节点中形成副本。

[0022] 特别地，所述步骤 A 具体包括：

[0023] A1、元数据节点通过节点监视器获取其它所有元数据节点的负载信息，并由负载最高的元数据节点计算出当前元数据集群的负载均衡度；当负载均衡度达到规定阈值时，执行步骤 A2；

[0024] A2、通过节点监视器获取其它所有元数据节点的负载信息和目录访问热度信息，目录热度监视器计算目录需要迁移的阈值以及目录子树迁移的目标元数据节点集合。

[0025] 特别地，所述步骤 B 具体包括：

[0026] B1、在目录子树迁移的目标元数据节点集合中，选取目录访问热度最低的元数据节点，查询该元数据节点的缓存中是否存在从目录访问热度最低的元数据节点迁入的目录，如果符合目录子树迁移条件，则将该类目录子树迁移到目录访问热度最低的元数据节点中；否则，在该元数据节点的目录中选择相应热度的目录子树迁移到目录访问热度最低的元数据节点中；

[0027] B2、如果没有目录子树达到迁移的条件，则从目标子树迁移的元数据节点集合中选择下一个元数据节点，重复执行步骤 B1。

[0028] 特别地，所述步骤 B1 具体包括：

[0029] B101、源元数据节点的目录迁移模块接收到目录子树迁移的消息后，冻结该目录子树，并向目标元数据节点发送需要迁移目录子树的路径消息；

[0030] B102、目标元数据节点获取需要迁移的目录子树到根目录路径上所有的目录节点的相关信息，创建缓存，并向源元数据节点发送确认信息；

[0031] B103、源元数据节点将要迁移的目录子树中的根节点以及其所有叶子结点中的目录节点数据打包发送到目标元数据节点；

[0032] B104、目标元数据节点将解包后的目录子树副本数据保存到步骤 B102 创建的缓存中；

[0033] B105、如果该被迁移的目录子树不存在副本，则跳转到步骤 B107；否则，向所有拥

有被迁移的目录子树副本的元数据节点发送目录子树迁移通知消息；

[0034] B106、所述拥有被迁移的目录子树副本的元数据节点修改被迁移目录子树的迁移状态，并发送确认消息；

[0035] B107、源元数据节点修改被迁移的目录子树的状态，并将目录子树中所有节点的导出信息打包发送到目标元数据节点中；

[0036] B108、目标元数据节点解包目录子树的导出信息，将其缓存到本地，并设置元数据信息的主副状态，再在副本集合中增加源元数据节点的记录，最后向源元数据节点发送确认信息；

[0037] B109、在源元数据节点上，如果该被迁移的目录子树不存在副本，则跳转到步骤B1011；否则，向所有拥有被迁移的目录子树副本的元数据节点发送导出通知信息，并修改迁移目录的迁移状态；

[0038] B1010、拥有被迁移的目录子树副本的元数据节点修改被迁移的目录子树的迁移状态，并向源元数据节点发送确认消息；

[0039] B1011、源元数据节点向目标元数据节点发送目录子树导出的完成消息，并解冻被迁移的目录子树；

[0040] B1012、目标元数据节点解冻被迁移的目录子树，迁移过程结束。

[0041] 特别地，所述步骤A2具体包括：

[0042] A201、源元数据节点根据元数据节点的负载信息，选择n个负载大小满足预设要求并且元数据服务处于正常状态的元数据节点作为目录子树的备份元数据节点，并向这些被备份节点发送目录子树复制的通知信息，然后将该目录子树的访问热度调整为原有热度的1/n；

[0043] A202、备份元数据节点向源元数据节点发送复制目录子树的查询消息；

[0044] A203、源元数据节点将被查询的目录子树的相关数据打包发送给备份元数据节点；

[0045] A204、备份元数据节点将目录子树解包，将目录子树的相关数据加载到本地缓存，修改复制目录的副本信息。

[0046] 本发明中元数据分布在元数据集群中的节点中，而不是内存容量受限的单一容量中，使得存储容量可横向扩展，并且元数据的分布使用动态分布策略，提高了分布式文件系统在企业应用环境所需的高可用性、存储容量以及访问性能的高可扩展性。

附图说明

[0047] 图1为本发明实施例提供的元数据管理系统拓扑结构示意图；

[0048] 图2为本发明实施例提供的元数据管理系统框图；

[0049] 图3为本发明实施例提供的元数据管理方法流程图。

具体实施方式

[0050] 下面结合附图和实施例对本发明作进一步说明。可以理解的是，此处所描述的具体实施例仅仅用于解释本发明，而非对本发明的限定。另外还需要说明的是，为了便于描述，附图中仅示出了与本发明相关的部分而非全部内容。

[0051] 请参照图 1 和图 2 所示，本实施例中分布式文件系统中的元数据管理系统由多台元数据服务器 101 组成，元数据服务器 101 之间通过网络通信定期交互彼此的节点信息，包括节点的健康信息、节点的负载量以及节点的访问热度。其中，所述元数据服务器 101 包括：消息引擎 201、节点监视器 202、元数据管理模块 203、元数据备份模块 204、元数据日志模块 205、目录热度监视器 206、目录复制模块 207 以及目录迁移模块 208。

[0052] 所述消息引擎 201 用于负责与文件系统客户端 2010 和存储节点 2011 之间通信。

[0053] 消息引擎 201 负责与文件系统客户端 2010 和存储节点 2011 之间通信，当文件系统客户端 2010 进行文件操作时，返回元数据信息，或者数据块映射信息，或者操作结果，并接收来自于存储节点 2011 定期上传的数据块映射信息。

[0054] 所述节点监视器 202 用于维护元数据集群中的每个节点的健康信息。所述健康信息具体包括元数据节点的工作负载、访问热度以及心跳情况。

[0055] 所述元数据管理模块 203 用于根据文件系统客户端 2010 的文件访问类型，完成对目录以及文件对像的操作请求。于本实施例，所述对目录以及文件对像的操作请求包括：新建目录、删除目录、读取目录、创建文件、删除文件以及重命名文件名称或者目录名称。

[0056] 所述元数据备份模块 204 用于根据元数据的备份要求，将相关的目录对象对应的元数据备份到元数据集群中的其它元数据节点的数据库 209 中。

[0057] 所述元数据日志模块 205 用于将文件系统客户端 2010 对文件系统操作的日志保存到本地。这样一来，当元数据节点发生故障异常退出时，可以根据保存的元数据日志恢复客户端对文件系统的操作，保持文件系统数据的一致性。

[0058] 所述目录热度监视器 206 用于实时监测目录对象访问的频繁程度，并根据其它元数据节点的负载情况，对目录对象是否需要复制和迁移进行决策。

[0059] 所述目录复制模块 207 用于根据目录热度监视器 206 发布的目录对象复制要求，完成目录对象向其它元数据节点的复制。

[0060] 所述目录迁移模块 208 用于根据目录热度监视器 206 发布的目录对象迁移要求，完成目录对象向其它元数据节点的迁移。

[0061] 基于上述元数据管理系统，本发明对应公开了一种元数据管理方法，如图 3 所示，该方法具体包括如下步骤：

[0062] 步骤 S301、根据元数据节点的负载，计算目录子树的访问频度，并根据该访问频度择出需要迁移的目录子树。

[0063] 元数据节点通过节点监视器获取其它所有元数据节点的负载信息，并由负载最高的元数据节点计算出当前元数据集群的负载均衡度；当负载均衡度达到规定阈值时，通过节点监视器获取其它所有元数据节点的负载信息和目录访问热度信息，目录热度监视器计算目录需要迁移的阈值以及目录子树迁移的目标元数据节点集合。

[0064] 步骤 S302、将需要迁移的目录子树迁移到负载较低的元数据节点上，达到元数据集群负载均衡。

[0065] 在目录子树迁移的目标元数据节点集合中，选取目录访问热度最低的元数据节点，查询该元数据节点的缓存中是否存在从目录访问热度最低的元数据节点迁入的目录，如果符合目录子树迁移条件，则将该类目录子树迁移到目录访问热度最低的元数据节点中；否则，在该元数据节点的目录中选择相应热度的目录子树迁移到目录访问热度最低的

元数据节点中。

[0066] 如果没有目录子树达到迁移的条件，则从目标子树迁移的元数据节点集合中选择下一个元数据节点，重复执行上述操作。

[0067] 步骤 S303、元数据节点根据目录的访问热度，将目录访问热度符合复制条件的目录子树复制到其它的元数据节点中形成副本。

[0068] 于本实施例，所述目录子树迁移的具体过程如下：

[0069] 一、源元数据节点的目录迁移模块接收到目录子树迁移的消息后，冻结该目录子树，并向目标元数据节点发送需要迁移目录子树的路径消息。

[0070] 二、目标元数据节点获取需要迁移的目录子树到根目录路径上所有的目录节点的相关信息，创建缓存，并向源元数据节点发送确认信息。

[0071] 三、源元数据节点将要迁移的目录子树中的根节点以及其所有叶子结点中的目录节点数据打包发送到目标元数据节点。

[0072] 四、目标元数据节点将解包后的目录子树副本数据保存到步骤二创建的缓存中。

[0073] 五、如果该被迁移的目录子树不存在副本，则跳转到步骤七；否则，向所有拥有被迁移的目录子树副本的元数据节点发送目录子树迁移通知消息。

[0074] 六、所述拥有被迁移的目录子树副本的元数据节点修改被迁移目录子树的迁移状态，并发送确认消息。

[0075] 七、源元数据节点修改被迁移的目录子树的状态，并将目录子树中所有节点的导出信息打包发送到目标元数据节点中。

[0076] 八、目标元数据节点解包目录子树的导出信息，将其缓存到本地，并设置元数据信息的主副状态，再在副本集合中增加源元数据节点的记录，最后向源元数据节点发送确认信息。

[0077] 九、在源元数据节点上，如果该被迁移的目录子树不存在副本，则跳转到步骤十一；否则，向所有拥有被迁移的目录子树副本的元数据节点发送导出通知信息，并修改迁移目录的迁移状态。

[0078] 十、拥有被迁移的目录子树副本的元数据节点修改被迁移的目录子树的迁移状态，并向源元数据节点发送确认消息。

[0079] 十一、源元数据节点向目标元数据节点发送目录子树导出的完成消息，并解冻被迁移的目录子树。

[0080] 十二、目标元数据节点解冻被迁移的目录子树，迁移过程结束。

[0081] 当元数据节点中的目录访问热度达到规定阈值时，启动目录子树的复制流程，其具体过程如下：

[0082] 一、源元数据节点根据元数据节点的负载信息，选择 n 个负载大小满足预设要求并且元数据服务处于正常状态的元数据节点作为目录子树的备份元数据节点，并向这些被备份节点发送目录子树复制的通知信息，然后将该目录子树的访问热度调整为原有热度的 1/n。其中，所述 n 为正整数，负载大小满足预设要求依据实际应用环境可灵活设定，本实施例中指负载较小。

[0083] 二、备份元数据节点向源元数据节点发送复制目录子树的查询消息。

[0084] 三、源元数据节点将被查询的目录子树的相关数据打包发送给备份元数据节点。

[0085] 四、备份元数据节点将目录子树解包，将目录子树的相关数据加载到本地缓存，修改复制目录的副本信息。

[0086] 本发明的技术方案提高了分布式文件系统在企业应用环境所需的高可用性、存储容量以及访问性能的高可扩展性。经过测试，使用本发明所述的元数据管理系统，分布式文件系统的平均性能提高了 2-4 倍，存储容量可以扩展到 30-60PB，支持的最大文件数量为 3 亿左右，系统平均修复时间为 30 分钟。

[0087] 注意，上述仅为本发明的较佳实施例及所运用技术原理。本领域技术人员会理解，本发明不限于这里所述的特定实施例，对本领域技术人员来说能够进行各种明显的变化、重复调整和替代而不会脱离本发明的保护范围。因此，虽然通过以上实施例对本发明进行了较为详细的说明，但是本发明不仅仅限于以上实施例，在不脱离本发明构思的情况下，还可以包括更多其他等效实施例，而本发明的范围由所附的权利要求范围决定。

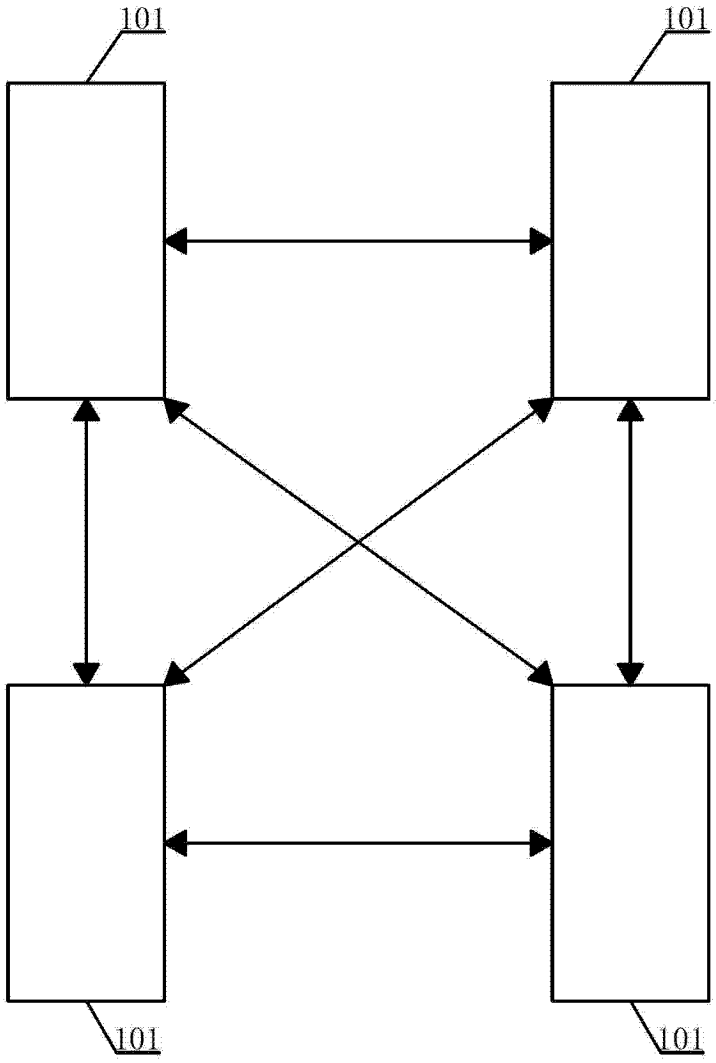


图 1

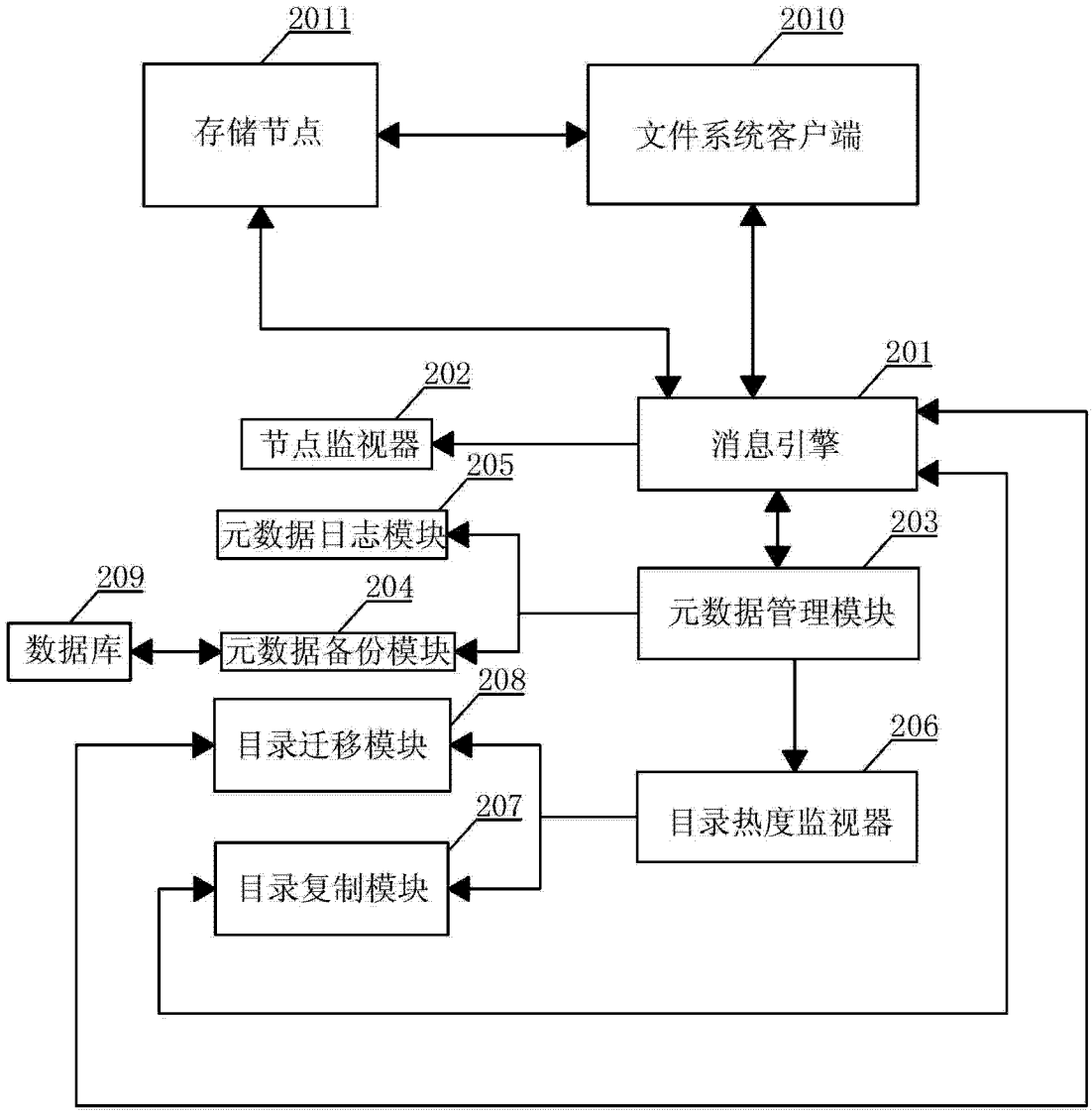

2011
存储节点
2010
文件系统客户端
201
消息引擎
202
节点监视器
205
元数据日志模块
203
元数据管理模块
209
数据库
204
元数据备份模块
208
目录迁移模块
206
目录热度监视器
207
目录复制模块

图 2

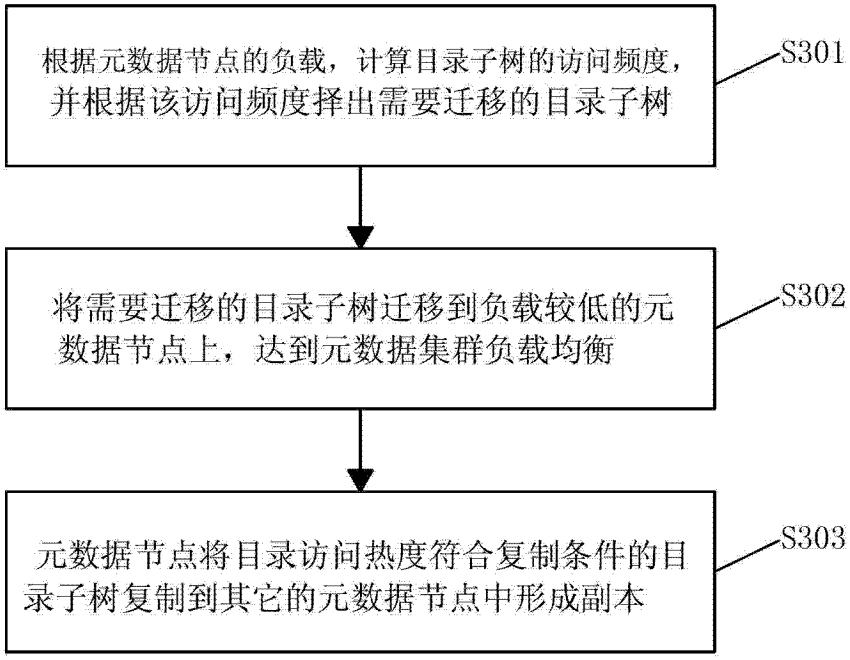

根据元数据节点的负载，计算目录子树的访问频度，并根据该访问频度择出需要迁移的目录子树
S301
将需要迁移的目录子树迁移到负载较低的元数据节点上，达到元数据集群负载均衡
S302
元数据节点将目录访问热度符合复制条件的目录子树复制到其它的元数据节点中形成副本
S303

图 3