(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 104615606 B
(45)授权公告日 2018.04.06

(21)申请号 201310544570.2

(22)申请日 2013.11.05

(65)同一申请的已公布的文献号
申请公布号 CN 104615606 A

(43)申请公布日 2015.05.13

(73)专利权人 阿里巴巴集团控股有限公司
地址 英属开曼群岛大开曼资本大厦一座四层847号邮箱

(72)发明人 刘磊

(74)专利代理机构 北京安信方达知识产权代理有限公司 11262
代理人 龙洪 栗若木

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

(56)对比文件
CN 103118133 A,2013.05.22,
CN 102946323 A,2013.02.27,
CN 103095769 A,2013.05.08,
CN 102693324 A,2012.09.26,
US 2012303579 A1,2012.11.29,

审查员 窦广健

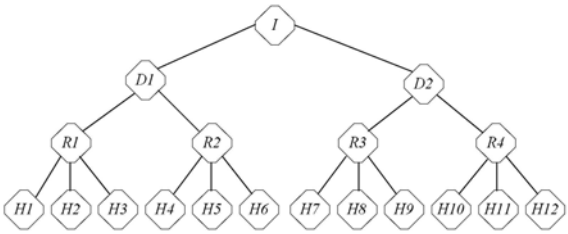
权利要求书2页 说明书9页 附图2页

(54)发明名称

一种Hadoop分布式文件系统及其管理方法

(57)摘要

一种Hadoop分布式文件系统及其管理方法，拓扑管理模块为Hadoop分布式文件系统创建树形网络拓扑结构，在代表集群的根节点和代表机架的第二层节点之间设置代表机房的第一层节点；副本配置模块在创建一跨机房文件时，指定多个机房来存放该文件的块副本，并为其中的每一机房配置存放的副本数；副本存放模块在所述多个机房存放所述块副本时，对其中的每一机房，按照为该机房配置的所述副本数，选择相同数目的数据节点存放所述块副本。采用上述Hadoop分布式文件系统及其管理方法，能够识别机房信息，在进行块副本存放、读取、维护等操作时，可根据机房的信息采用合理的策略，节约跨机房的带宽资源，提高系统性能。



CN 104615606 B

1.一种Hadoop分布式文件系统跨机房的管理方法,包括:

Hadoop分布式文件系统创建树形网络拓扑结构,在代表集群的根节点和代表机架的第二层节点之间设置代表机房的第一层节点;

创建一跨机房文件时,指定多个机房来存放该文件的块副本,并为其中的每一机房配置存放的副本数;

在所述多个机房存放所述块副本时,对其中的每一机房,按照为该机房配置的所述副本数,选择相同数目的数据节点存放所述块副本。

2.如权利要求1所述的方法,其特征在于,还包括:

对所述跨机房文件的块副本进行维护时,先确定存放有所述块副本的所有机房和其中每一机房的实际存放数;对每一机房,如实际存放数不等于为该机房配置的所述副本数,则在该机房内对所述块副本进行复制或删除以使实际存放数等于配置的所述副本数。

3.如权利要求2所述的方法,其特征在于:

指定多个机房来存放该文件的块副本时,优先指定用户所在的机房;

对所述块副本进行复制时,优先将同一个机房内存放有所述块副本的数据节点作为源。

4.如权利要求1或2或3所述的方法,其特征在于,还包括:

接收到用户读取块副本的指令,选择读取的所述块副本时,优先选择用户所在数据节点存放的所述块副本,其次选择用户所在机架存放的所述块副本,再次选择用户所在机房存放的所述块副本,最后选择其他机房存放的所述块副本。

5.如权利要求1或2或3所述的方法,其特征在于,还包括:

使用平衡工具平衡Hadoop分布式文件系统集群数据节点的磁盘利用率时,只在一个机房内的各数据节点之间进行平衡。

6.如权利要求1或2或3所述的方法,其特征在于:

所述指定多个机房来存放该文件的块副本,包括:

配置该文件的文件路径与多个机房的对应关系,使用所述文件路径对应的多个机房存放该文件的块副本。

7.一种跨机房的Hadoop分布式文件系统,包括:

拓扑管理模块,用于创建Hadoop分布式文件系统HDFS的树形网络拓扑结构,在代表集群的根节点和代表机架的第二层节点之间增加代表机房的第一层节点;

副本配置模块,用于在创建跨机房文件时,指定多个机房来存放该文件的块副本,并为其中的每一机房配置存放的副本数;

副本存放模块,用于在多个机房存放跨机房文件的副本块时,对其中的每一机房,按照为该机房配置的副本数,选择相同数目的数据节点存放所述块副本。

8.如权利要求7所述的系统,其特征在于,还包括:

副本维护模块,用于对所述跨机房文件的块副本进行维护时,先确定存放有所述块副本的所有机房和其中每一机房的实际存放数;对每一机房,如实际存放数不等于为该机房配置的所述副本数,则在该机房内对所述块副本进行复制或删除以使实际存放数等于配置的所述副本数。

9.如权利要求8所述的系统,其特征在于:

所述副本配置模块指定多个机房来存放该文件的块副本时，优先指定用户所在的机房；

所述副本维护模块对所述块副本进行复制时，优先将同一个机房内存放有所述块副本的数据节点作为源。

10.如权利要求7或8或9所述的系统，其特征在于，还包括：

数据读取模块，用于在接收到用户读取块副本的指令，对块副本进行读取时，优先选择用户所在数据节点的块副本，其次选择用户所在机架的块副本，再次选择用户所在机房的块副本，最后选拔其他机房的块副本。

11.如权利要求7或8或9所述的系统，其特征在于，还包括：

性能优化模块，用于在使用平衡工具平衡Hadoop分布式文件系统集群的数据节点的磁盘利用率时，只在一个机房内的各数据节点之间进行平衡。

12.如权利要求7或8或9所述的系统，其特征在于：

所述副本配置模块指定多个机房来存放该文件的块副本，包括：配置该文件的文件路径对应的多个机房，使用所述文件路径对应的多个机房存放该文件的块副本。

一种Hadoop分布式文件系统及其管理方法

技术领域

[0001] 本申请涉及Hadoop分布式文件系统(HDFS,Hadoop Distributed File System),更具体地,涉及一种跨机房的Hadoop分布式文件系统及相应的管理方法。

背景技术

[0002] Hadoop是Internet上对搜索关键字进行内容分类的工具。Hadoop由Apache Software Foundation公司于2005年秋天作为Lucene的子项目Nutch的一部分正式引入。Hadoop分布式文件系统被设计成适合运行在通用硬件(commodity hardware)上的分布式文件系统。HDFS是一个高度容错性(fault-tolerant)的系统,适合部署在廉价(low-cost)的机器上。HDFS能提供高吞吐量的数据访问,非常适合大规模数据集(large data set)上的应用。

[0003] HDFS用于存储超大的文件,文件内容被分解成多个块(block),每个block默认为64M。为了提供高可靠性,一个block的内容会被复制成多份,存储在不同的物理机器上。一个HDFS集群是由一个名字节点(NameNode)和多个数据节点(DataNodes)组成。NameNode是一个中心服务器,负责管理文件系统的名字空间(namespace)以及客户端对文件的访问,是所有HDFS元数据的仲裁者和管理者。DataNode用于存储块副本,并提供对块副本的读取等操作。

[0004] HDFS会创建一个如图1所示的网络拓扑结构,根据拓扑结构来选择存放副本的DataNode。图中,根节点I代表整个HDFS集群,第一层节点R1～R4代表机架,叶子节点H1～H12代表DataNode。随着HDFS集群规模的不断扩大,一个机房内物理机器的数量无法满足集群规模的需要,此时需要把一个文件的块副本存储在多个机房内。但目前HDFS创建的网络拓扑结构无法获得任何机房信息,不能取得令人满意的性能。

[0005] 申请内容

[0006] 本申请要解决的技术问题是提供一种Hadoop分布式文件系统及其管理方法,能够基于机房信息有效管理文件,提高系统性能。

[0007] 为了解决上述问题,本申请提供了一种Hadoop分布式文件系统跨机房的管理方法,包括:

[0008] Hadoop分布式文件系统创建树形网络拓扑结构,在代表集群的根节点和代表机架的第二层节点之间设置代表机房的第一层节点;

[0009] 创建一跨机房文件时,指定多个机房来存放该文件的块副本,并为其中的每一机房配置存放的副本数;

[0010] 在所述多个机房存放所述块副本时,对其中的每一机房,按照为该机房配置的所述副本数,选择相同数目的数据节点存放所述块副本。

[0011] 较佳地,上述方法还包括:

[0012] 对所述跨机房文件的块副本进行维护时,先确定存放有所述块副本的所有机房和其中每一机房的实际存放数;对每一机房,如实际存放数不等于为该机房配置的所述副本

数，则在该机房内对所述块副本进行复制或删除以使实际存放数等于配置的所述副本数。

[0013] 较佳地，

[0014] 指定多个机房来存放该文件的块副本时，优先指定用户所在的机房；

[0015] 对所述块副本进行复制时，优先将同一个机房内存放有所述块副本的数据节点作为源。

[0016] 较佳地，上述方法还包括：

[0017] 接收到用户读取块副本的指令，选择读取的所述块副本时，优先选择用户所在数据节点存放的所述块副本，其次选择用户所在机架存放的所述块副本，再次选择用户所在机房存放的所述块副本，最后选择其他机房存放的所述块副本。

[0018] 较佳地，上述方法还包括：

[0019] 使用平衡工具平衡Hadoop分布式文件系统集群数据节点的磁盘利用率时，只在一个机房内的各数据节点之间进行平衡。

[0020] 较佳地，

[0021] 所述指定多个机房来存放该文件的块副本，包括：

[0022] 配置该文件的文件路径与多个机房的对应关系，使用所述文件路径对应的多个机房存放该文件的块副本。

[0023] 相应地，本申请提供的跨机房的Hadoop分布式文件系统，包括：

[0024] 拓扑管理模块，用于创建Hadoop分布式文件系统HDFS的树形网络拓扑结构，在代表集群的根节点和代表机架的第二层节点之间增加代表机房的第一层节点；

[0025] 副本配置模块，用于在创建跨机房文件时，指定多个机房来存放该文件的块副本，并为其中的每一机房配置存放的副本数；

[0026] 副本存放模块，用于在多个机房存放跨机房文件的副本块时，对其中的每一机房，按照为该机房配置的副本数，选择相同数目的数据节点存放所述块副本。

[0027] 较佳地，上述系统还包括：

[0028] 副本维护模块，用于对所述跨机房文件的块副本进行维护时，先确定存放有所述块副本的所有机房和其中每一机房的实际存放数；对每一机房，如实际存放数不等于为该机房配置的所述副本数，则在该机房内对所述块副本进行复制或删除以使实际存放数等于配置的所述副本数。

[0029] 较佳地，

[0030] 所述副本配置模块指定多个机房来存放该文件的块副本时，优先指定用户所在的机房；

[0031] 所述副本维护模块对所述块副本进行复制时，优先将同一个机房内存放有所述块副本的数据节点作为源。

[0032] 较佳地，上述系统还包括：

[0033] 数据读取模块，用于在接收到用户读取块副本的指令，对块副本进行读取时，优先选择用户所在数据节点的块副本，其次选择用户所在机架的块副本，再次选择用户所在机房的块副本，最后选拔其他机房的块副本。

[0034] 较佳地，上述系统还包括：

[0035] 性能优化模块，用于在使用平衡工具平衡Hadoop分布式文件系统集群的数据节点

的磁盘利用率时，只在一个机房内的各数据节点之间进行平衡。

[0036] 较佳地，

[0037] 所述副本配置模块指定多个机房来存放该文件的块副本，包括：配置该文件的文件路径对应的多个机房，使用所述文件路径对应的多个机房存放该文件的块副本。

[0038] 采用上述Hadoop分布式文件系统及其管理方法，能够识别机房信息，在进行块副本存放、读取、维护等操作时，可根据机房的信息采用合理的策略，节约跨机房的带宽资源，提高系统性能。

附图说明

[0039] 图1是现有HDFS网络拓扑结构图；

[0040] 图2是本申请实施例一HDFS跨机房的管理方法的流程图；

[0041] 图3是本申请实施例一包含机房信息的HDFS网络拓扑结构图；

[0042] 图4是本申请实施例一Hadoop分布式文件系统的模块图；

[0043] 图5是本申请实施例二对跨机房文件的块副本进行维护的流程图。

具体实施方式

[0044] 为使本申请的目的、技术方案和优点更加清楚明白，下文中将结合附图对本申请的实施例进行详细说明。需要说明的是，在不冲突的情况下，本申请中的实施例及实施例中的特征可以相互任意组合。

[0045] 在本申请一个典型的配置中，HDFS系统的各个节点包括一个或多个处理器(CPU)、输入/输出接口、网络接口和内存。

[0046] 内存可能包括计算机可读介质中的非永久性存储器，随机存取存储器(RAM)和/或非易失性内存等形式，如只读存储器(ROM)或闪存(flashRAM)。内存是计算机可读介质的示例。

[0047] 计算机可读介质包括永久性和非永久性、可移动和非可移动媒体可以由任何方法或技术来实现信息存储。信息可以是计算机可读指令、数据结构、程序的模块或其他数据。计算机的存储介质的例子包括，但不限于相变内存(PRAM)、静态随机存取存储器(SRAM)、动态随机存取存储器(DRAM)、其他类型的随机存取存储器(RAM)、只读存储器(ROM)、电可擦除可编程只读存储器(EEPROM)、快闪记忆体或其他内存技术、只读光盘只读存储器(CD-ROM)、数字多功能光盘(DVD)或其他光学存储、磁盒式磁带，磁带磁磁盘存储或其他磁性存储设备或任何其他非传输介质，可用于存储可以被计算设备访问的信息。按照本文中的界定，计算机可读介质不包括非暂存电脑可读媒体(transitory media)，如调制的数据信号和载波。

[0048] 实施例一

[0049] HDFS集群将一个跨机房文件的块副本存放到不同机房的数据节点，有利于提高数据的可靠性，避免因某一机房发生意外故障而造成数据无法读取。在需要把该文件的块副本存储在多个机房内的情况下，由于机房间的带宽有限而且十分昂贵，存放时要尽量减少跨机房的数据读取以减少跨机房的带宽占用。但是，使用现有的HDFS网络拓扑结构树无法获得任何机房的信息，导致所有的块副本有可能被放置在同一个机房的数据节点内，无法实现跨机房的合理存放。

[0050] 本实施例HDFS跨机房的管理方法如图2所示，包括：

[0051] 步骤110，Hadoop分布式文件系统创建树形网络拓扑结构时，在代表集群的根节点和代表机架的第二层节点之间增加代表机房的第一层节点；

[0052] 步骤120，创建跨机房文件时，指定多个机房来存放该文件的块副本，为其中的每一机房配置存放的副本数；

[0053] 步骤130，存放所述块副本时，对所述多个机房中的每一机房，按照为该机房配置的所述副本数，选择相同数目的数据节点存放所述块副本。

[0054] 其中，

[0055] 在步骤110中，创建的HDFS网络拓扑结构如图3所示，根节点I代表整个集群，第一层节点D1～D2代表机房，第二层节点R1～R4代表机架，叶子节点H1～H12代表DataNode。可见，新的网络拓扑结构仍是一种树形网络拓扑结构，但在代表集群的根节点和代表机架的第二层节点之间增加了代表机房的第一层节点。根据新的网络拓扑结构，可以识别出有哪些机房，并且识别出每个机房包含哪些机架和DataNode。

[0056] 在步骤120中，

[0057] 当用户创建一个跨机房文件时，文件包含的块副本要放置在多个机房内。选择块副本存放位置时，较佳地，如果用户在某个机房内，在指定多个机房来存放该文件的块副本时，优先指定用户所在的机房，也就是说，除非用户所在机房无法存放，否则指定的多个机房中要包含用户所在的机房。这样可以把块副本的数据尽量写在用户所在的机房，减少跨机房的网络带宽占用，并且提升写的速度。

[0058] 指定多个机房来存放该文件的块副本，具体地，可以是配置该文件的文件路径与多个机房的对应关系，使用所述文件路径对应的多个机房存放该文件的块副本。基于文件路径(文件路径可以是目录也可以是文件)来指定机房非常灵活，便于修改。用户可以根据业务要求和硬件资源的使用情况，决定选择哪些文件跨机机房存放，哪些文件不跨机房存放。

[0059] 对其中的每一机房配置存放的副本数时，对主机房(NameNode所在的机房)存放的副本数，可以在创建该文件或调用DFS客户端设置副本的方法如DFSClient.setReplication来指定，其他机房存放的副本数可以通过配置文件指定。

[0060] 例如：NameNode属于的主机房记为d1，用户创建“/user/people.txt”文件时指定在主机房d1存放的副本数为3。针对其他机房的配置文件的内容为“/user/people.txt,d2:d3,2:4”，表示/user/people.txt文件的副本存放在d2和d3机房，d2机房存放2个副本，d3机房包含4个副本。

[0061] 对上述配置信息可以随时更新，即随时修改文件路径的跨机房信息，如：

[0062] 4月10号的配置文件为：

[0063] /group1/table1/2013-04-10 d2:d3,2:4

[0064] /group1/table2/2013-04-10 d2:d3,2:4

[0065] 其中，/group1/table1/2013-04-10及/group1/table2/2013-04-10表示文件路径。

[0066] 4月11号对配置内容更新如下：

[0067] /group1/table2/2013-04-10 d2:d3,3:3

[0068] /group3/table0/2013-04-10 d2:d3,3:3

[0069] 则最新的内容为：

[0070] /group1/table1/2013-04-10 d2:d3,2:4 del

[0071] /group1/table2/2013-04-10 d2:d3,3:3 up

[0072] /group3/table0/2013-04-10 d2:d3,3:3 add

[0073] 配置的文件路径对应的跨机房信息的变化，会改变相应文件的块副本在多个机房内的分布。

[0074] 在步骤130中，

[0075] 在每个机房内根据机架信息存放副本时，采用以下的存放策略：

[0076] •选择存放第一个副本的数据节点

[0077] 如果用户不在一个DataNode上，则在本机房内随机选择一个机上的DataNode存放第一个副本。

[0078] 如果client在一个DataNode上，则选择这个DataNode存放第一个副本。

[0079] •选择存放第二副本的DataNode

[0080] 存放第二个副本的DataNode，与存放第一个副本的DataNode在同一个机房内但不在同一个机架上。

[0081] •选择存放第三个副本的DataNode

[0082] 存放第三个副本的DataNode，与存放第二个副本的DataNode在同一个机房的同一个机架上。

[0083] •选存放第四个及更多的副本

[0084] 在本机房内随机选择机架存放副本。

[0085] •约束条件

[0086] 确保一个DataNode不会存放一个以上的副本。

[0087] 如果副本的个数小于1/2机架总数，确保一个机房内的一个机架不会保存两个以上的副本。

[0088] 基于上述新的网络拓扑结构树，可以在HDFS中新增如下的应用程序编程接口(Application Programming Interface，API)以感知机房信息：

[0089] •public int getNumOfRacks(String datacenter)

[0090] 该API用于获得机房包含的机架个数

[0091] •public boolean contains(String datacenter,Node node)

[0092] 该API用于判断一个机房是否包含一个节点(node，node可以是机房、机架或者datanode)

[0093] •public int getNumOfLeaves(String datacenter)

[0094] 该API用于获得一个机房包含了多少个DataNode

[0095] •public String getDataCenter(Node node)

[0096] 该API用于获得DataNode所在的机房的名称

[0097] •public boolean isOnSameDatacenter(Node node1,Node node2)

[0098] 该API用于检测两个DataNode是否在同一个机房

[0099] 另外，可以在一些已有的API中将机房作为新增参数，例如：

[0100] • public int countNumOfAvailableNodes(String scope,

[0101] String excludedScope,

[0102] Collection<Node>excludedNodes)

[0103] 该API用于获得在scope内但不在excludedScope和excludedNodes中的DataNode节点的个数。基于新的网络拓扑结构树，可以将scope指定为机房d1，excludedScope指定为机房d1中的机架rack1，excludedNodes表示scope范围内块副本不应存放的DataNode。

[0104] • public Node chooseRandom(String scope,String excludedScope)

[0105] 该API用于选择在scope范围内但不在excludedScope中的一个DataNode，其中的参数scope、excludedScope均可以为机房。

[0106] • public void pseudoSortByDistance(Node reader,Node[]nodes)

[0107] 该API用于根据读取者(reader)在local node、local rack与local datacenter对nodes数组进行排序，其中，local node表示：reader和datanode在同一个节点上。local rack表示：reader和datanode在同一个机架上。local datacenter表示：reader和datanode在同一个机房内。local datacenter为新增参数。

[0108] 相应地，本实施例还提供了一种跨机房的Hadoop分布式文件系统，如图4所示，包括：

[0109] 拓扑管理模块11，用于在创建HDFS的树形网络拓扑结构时，在代表集群的根节点和代表机架的第二层节点之间增加代表机房的第一层节点。

[0110] 副本配置模块12，用于在创建跨机房文件时，配置多个机房存放该文件的块副本，并为其中的每一机房配置存放的副本数。

[0111] 副本存放模块13，用于在多个机房存放跨机房文件时，对其中的每一机房，按照为该机房配置的所述副本数，选择相同数目的数据节点存放所述块副本。

[0112] 较佳地，副本配置模块12指定多个机房来存放该文件的块副本，包括：配置该文件的文件路径对应的多个机房，使用所述文件路径对应的多个机房存放该文件的块副本。

[0113] 较佳地，副本配置模块12指定多个机房来存放该文件的块副本时，优先指定用户所在的机房；

[0114] 较佳地，副本配置模块12为其中的每一机房配置存放的副本数，包括：在创建该文件时或调用DFSClient.setReplication时指定主机房存放的副本数，通过配置文件指定其他机房存放的副本数。

[0115] 可选地，本实施例的Hadoop分布式文件系统还可以包括：

[0116] 编程接口模块，用于基于所述树形网络拓扑结构，在已有应用程序编程接口API中将机房作为新增参数，并增加以下API中的一种或多种以感知机房信息：

[0117] 获得机房包含的机架个数的API；

[0118] 判断一个机房是否包含某一节点的API；

[0119] 获得一个机房包含了多少个数据节点的API；

[0120] 获得数据节点所在的机房的名称的API；

[0121] 检测两个数据节点是否在同一个机房的API。

[0122] 本实施例在创建HDFS网络拓扑结构树时，增加了代表机房的一层节点，可以根据可靠性、节约带宽资源等策略，选择合适的机房配置跨机房文件的副本数并和存放块副本。

[0123] 实施例二

[0124] 在HDFS运行过程中，存放副本的DataNode可能死掉，死掉后又可能重启，因而块副本的个数可能会小于或多于配置的要求。如块副本个数小于要求的个数，block为under状态，如块副本个数大于要求的个数，block为over状态。对块副本维护时，当实际存放的块副本的个数小于配置的副本数时要复制块副本，当实际存放的块副本的个数大于配置的副本数时要将多出的块副本进行删除。

[0125] 已有HDFS方案中判断block是否为under或over状态时，是根据集群中块副本的总个数来判断的，不能满足机房内副本数的配置要求。

[0126] 基于实施例一的网络拓扑结构树和对跨机房文件的块副本进行配置、存放的方法，本实施例提供了一种对所述跨机房文件的块副本进行维护的方法，如图5所示，包括：

[0127] 步骤210，确定存放有所述块副本的所有机房和其中每一机房的实际存放数；

[0128] 存放块副本的数据节点和所属的机房可以根据存放时的记录来确定。

[0129] 步骤220，对每一机房，如实际存放数不等于为该机房配置的所述副本数，在该机房内对所述块副本进行复制或删除，使实际存放数等于配置的所述副本数。

[0130] 为机房配置的所述副本数见实施例一中的说明。本步骤中，在机房内对所述块副本进行复制时，为了减少跨机房的带宽的占用，较佳选择同一个机房内存放有所述块副本的DataNode作为源进行复制工作。

[0131] 下面通过一个示例进行说明：

[0132] 假定，配置内容为”/group/user.txt dc2:3”，即机房dc2存放该文件的3个块副本，另外，为主机房dc1配置的块副本个数为3。则期望的块副本总个数为6，维护时要分别判断每个机房内存放的块副本个数是否达到配置的要求。

[0133] 如机房dc1和dc2都实际存放有3个块副本，并且存储的块副本总数也为6，则block的块副本分布是满足要求的。

[0134] 如dc1中实际存放的块副本个数为3，dc2中实际存放的块副本个数小于3，则block为under状态，需要在dc2中复制一个新的块副本。

[0135] 如dc1中实际存放的块副本个数大于3，dc2中实际存放的块副本个数为3，则block为over状态，需要对dc1机房内多余的副本进行删除。

[0136] 如dc1中实际存放的块副本个数为4，dc2中实际存放的块副本个数为2，虽然块副本总数为6，但dc1多存放了一个副本，dc2少存放了一个副本，则该block即是over状态也是under状态，需要从dc1中删除一个块副本，并且在dc2中复制一个新的块副本。

[0137] 相应地，本实施例HDFS系统在实施例一包含的模块的基础上，还包括：

[0138] 副本维护模块，用于对所述跨机房文件的块副本进行维护时，先确定存放有所述块副本的所有机房和其中每一机房的实际存放数；对每一机房，如实际存放数不等于为该机房配置的所述副本数，则在该机房内对所述块副本进行复制或删除以使实际存放数等于配置的所述副本数。较佳地，对所述块副本进行复制时，优先将同一个机房内存放有所述块副本的数据节点作为源。

[0139] 实施例三

[0140] 本实施例在实施例一的基础上，提供了一种用户读取数据的方法，要尽量选择与用户同一个机房内的块副本进行数据读取，以便减少跨机房网络带宽的占用。这个过程需

要考虑存放块副本的DataNode与用户之间的距离，选择一个离用户最近的DataNode进行数据读取。

[0141] 选取DataNode的顺序如下：

[0142] 如果用户所在DataNode存放有要读取的块副本，则选择本地的DataNode；

[0143] 如果用户所在机架内的DataNode存放有所述块副本，则随机选择这个机架内存放有所述块副本的一个DataNode。

[0144] 如果用户所在机房内的DataNode存放有所述块副本，则随机选择这个机房内存放有所述块副本的一个DataNode。

[0145] 如果用户不在任何机房内，则从其他机房内存放有所述块副本的DataNode中随机选择一个DataNode。

[0146] 也就是说，接收到用户读取块副本的指令，选择读取的所述块副本时，优先选择用户所在数据节点存放的所述块副本，其次选择用户所在机架存放的所述块副本，再次选择用户所在机房存放的所述块副本，最后选择其他机房存放的所述块副本。

[0147] 相应地，本实施例提供的HDFS系统在实施例一包含的模块的基础上，还包括：

[0148] 数据读取模块，用于在接收到用户读取块副本的指令，选择读取的所述块副本时，优先选择用户所在数据节点存放的所述块副本，其次选择用户所在机架存放的所述块副本，再次选择用户所在机房存放的所述块副本，最后选拔其他机房存放的所述块副本。

[0149] 实施例四

[0150] 本实施例在实施例一的基础上，提供了一种HDFS系统中的平衡(Balancer)方法，Balancer是一个hadoop的平衡工具，用于平衡HDFS集群的DataNode的磁盘利用率。现有的Balancer方法并没有考虑机房信息，这会导致副本的分布不符合跨机房的分布。需要修改这个工具使Balancer感知机房，并且只在一个机房内进行平衡。

[0151] 本实施例提供一种HDFS的平衡方法，在使用平衡工具平衡HDFS集群的DataNode的磁盘利用率时，只在一个机房内的各数据节点之间进行平衡。

[0152] 例如:有d1和d2两个机房

[0153] 命令“./bin/start-balancer.sh d1“只对d1机房内的所有DataNode进行平衡。

[0154] 命令“./bin/start-balancer.sh d2“只对d2机房内的所有DataNode进行平衡。

[0155] 相应地，本实施例提供的HDFS系统在实施例一包含的模块的基础上，还包括：

[0156] 性能优化模块，用于在使用平衡工具平衡HDFS集群的DataNode的磁盘利用率时，只在一个机房内的各数据节点之间进行平衡。

[0157] 对于本申请的HDFS系统，上述实施例二的副本维护模块、实施例三的数据读取模块及实施例四的性能优化模块可以任意组合。

[0158] 本领域普通技术人员可以理解上述方法中的全部或部分步骤可通过程序来指令相关硬件完成，所述程序可以存储于计算机可读存储介质中，如只读存储器、磁盘或光盘等。可选地，上述实施例的全部或部分步骤也可以使用一个或多个集成电路来实现，相应地，上述实施例中的各模块/单元可以采用硬件的形式实现，也可以采用软件功能模块的形式实现。本申请不限制于任何特定形式的硬件和软件的结合。

[0159] 以上所述仅为本申请的优选实施例而已，并不用于限制本申请，对于本领域的技术人员来说，本申请可以有各种更改和变化。凡在本申请的精神和原则之内，所作的任何修

改、等同替换、改进等，均应包含在本申请的保护范围之内。

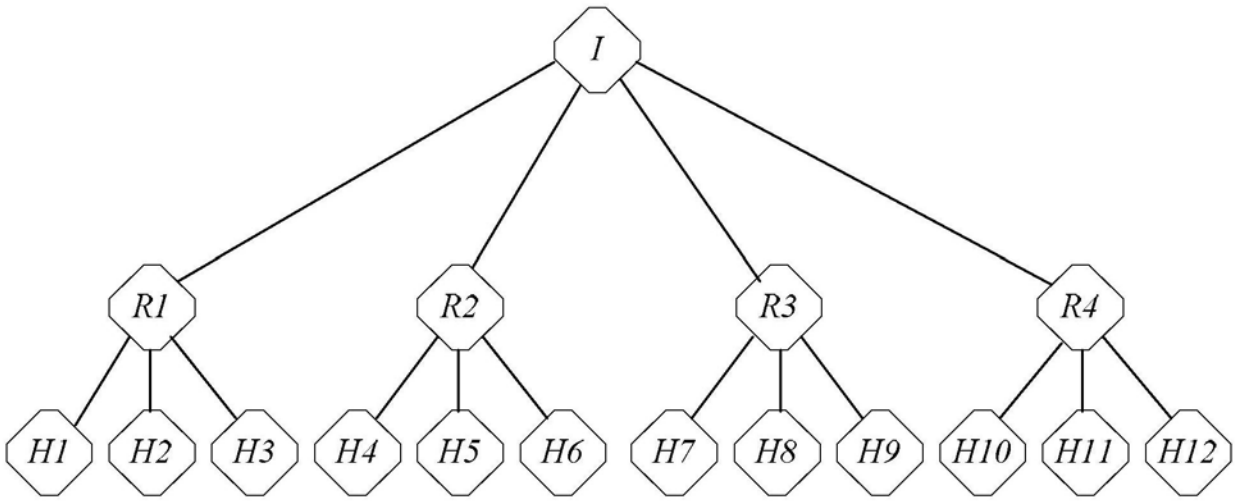

I
R1
R2
R3
R4
H1
H2
H3
H4
H5
H6
H7
H8
H9
H10
H11
H12

图1

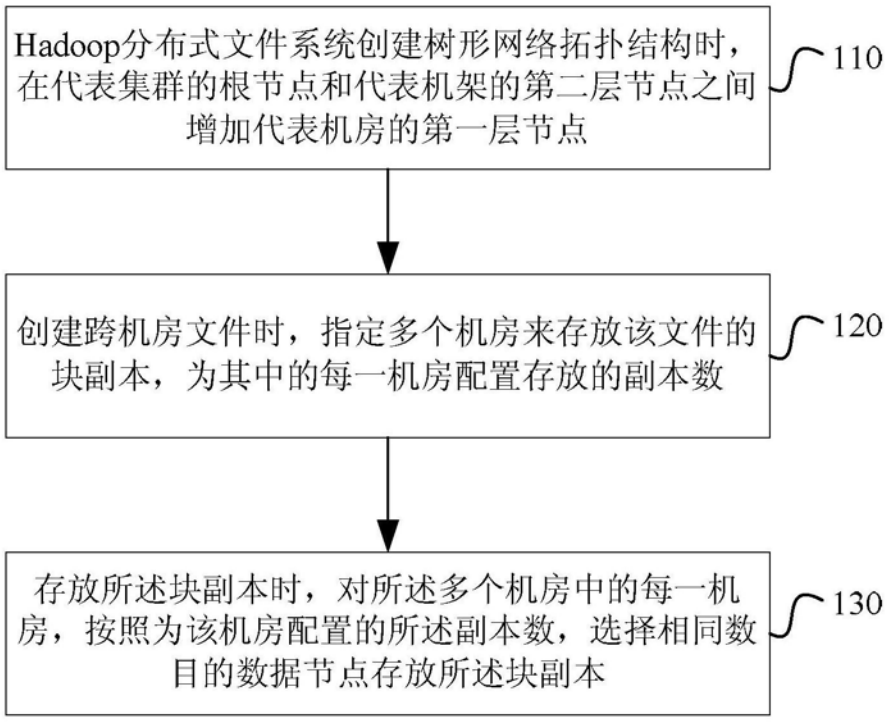

Hadoop分布式文件系统创建树形网络拓扑结构时，在代表集群的根节点和代表机架的第二层节点之间增加代表机房的第一层节点
110
创建跨机房文件时，指定多个机房来存放该文件的块副本，为其中的每一机房配置存放的副本数
120
存放所述块副本时，对所述多个机房中的每一机房，按照为该机房配置的所述副本数，选择相同数目的数据节点存放所述块副本
130

图2

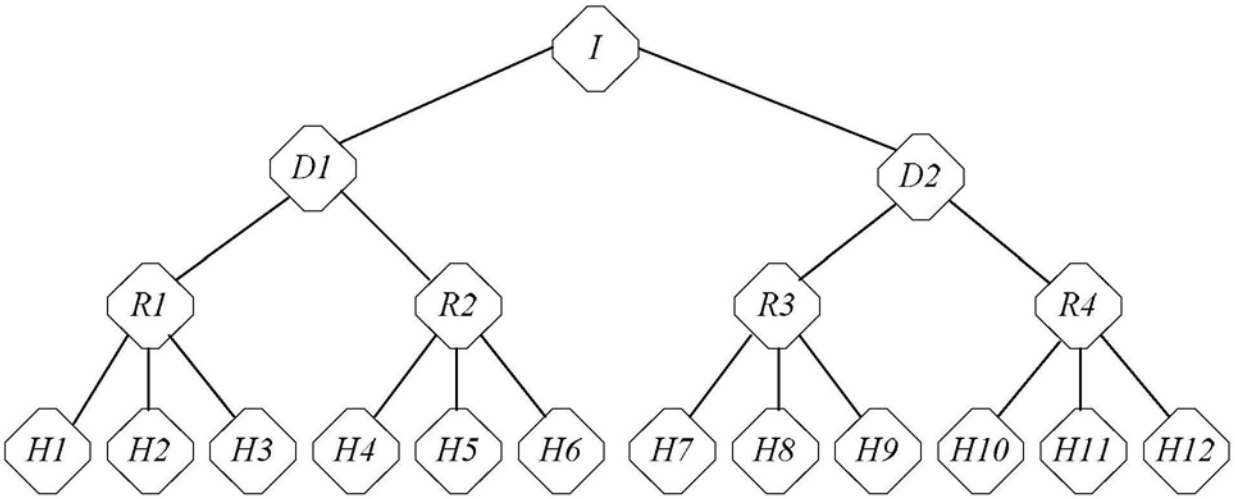

I
D1
D2
R1
R2
R3
R4
H1
H2
H3
H4
H5
H6
H7
H8
H9
H10
H11
H12

图3

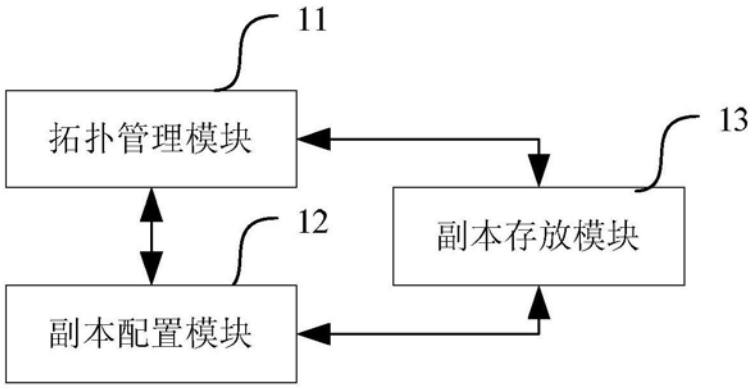

11
拓扑管理模块
13
副本存放模块
12
副本配置模块

图4

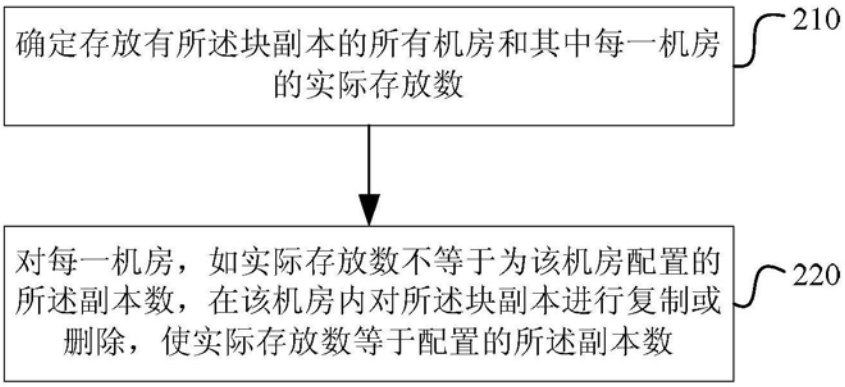

确定存放有所述块副本的所有机房和其中每一机房的实际存放数
210
对每一机房，如实际存放数不等于为该机房配置的所述副本数，在该机房内对所述块副本进行复制或删除，使实际存放数等于配置的所述副本数
220

图5