(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 104915376 A
(43)申请公布日 2015.09.16

(21)申请号 201510223848.5

(22)申请日 2015.05.05

(71)申请人 华南理工大学
地址 510640 广东省广州市天河区五山路381号

(72)发明人 李磊 李达港 金连文

(74)专利代理机构 广州市华学知识产权代理有限公司 44245
代理人 罗观祥

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

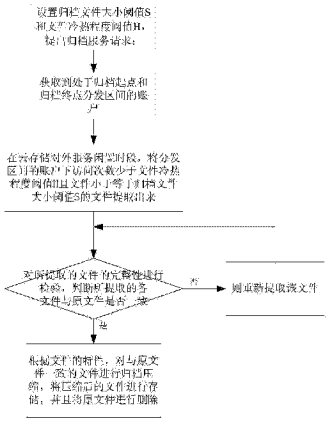
权利要求书1页 说明书4页 附图1页

(54)发明名称

一种云存储中文件的归档压缩方法

(57)摘要

本发明公开了一种云存储中文件的归档压缩方法，步骤：设置归档文件大小阈值S和文件冷热程度阀值H；根据归档服务请求分发归档起点和归档终点，获取到处于归档起点和归档终点分发区间的账户；在云存储对外服务闲置时段，将各账户下访问次数少于文件冷热程度阀值H且文件小于等于归档文件大小阈值S的文件提取出来；对所提取的文件的完整性进行检验，判断所提取的文件与原文件是否一致；若否，则重新提取该文件，针对该文件的完整性进行重新检验，直到获取到与原文件相同的文件；根据文件的特性，对与原文件一致的文件进行归档压缩及存储。本发明根据云存储中文件访问热度进行归档压缩，实现了云存储文件数目增加速度收敛和存储效益的提高。



CN 104915376 A

1.一种云存储中文件的归档压缩方法，其特征在于，步骤如下：

S1、设置归档文件大小阈值 S 和文件冷热程度阀值 H，通过归档服务进程提出归档服务请求；

S2、根据归档服务进程的归档服务请求分发归档起点和归档终点，然后获取到处于归档起点和归档终点分发区间的账户；

S3、在云存储对外服务闲置时段，归档服务进程执行任务：归档服务进程依次遍历处于归档起点和归档终点分发区间的账户，将各账户下访问次数少于文件冷热程度阀值 H 且文件小于等于归档文件大小阈值 S 的文件提取出来；

S4、对所提取的文件的完整性进行检验，判断所提取的各文件与原文件是否一致；

若否，则重新提取该文件，然后针对该文件的完整性进行重新检验，直到获取到与原文件相同的文件；

若是，则进入步骤 S5；

S5、根据文件的特性，对步骤 S4 中获取的与原文件一致的文件进行归档压缩，然后将压缩后的文件存储到云存储中，并且将云存储中对应的原文件进行删除。

2.根据权利要求 1 所述的云存储中文件的归档压缩方法，其特征在于，还包括以下步骤：

设置访问时间间隔 I，当文件被访问时，判断该文件是否已归档压缩；

若是，则查询文件的具体存储路径，然后从压缩文件中提取出目标文件并返回文件的内容；

若否，则判断该文件当前访问时间与上次访问时间之差是否超过访问时间间隔 I，若是，则将该文件的访问次数置 1，若否，则将其访问次数加 1。

3.根据权利要求 1 所述的云存储中文件的归档压缩方法，其特征在于，所述访问时间间隔 I 为 15 天以上。

4.根据权利要求 1 所述的云存储中文件的归档压缩方法，其特征在于，还包括以下步骤：文件写入时，将其访问次数置为文件冷热程度阀值 H。

5.根据权利要求 1 所述的云存储中文件的归档压缩方法，其特征在于，所述步骤 S2 中，归档起点至归档终点分发区间的账户是按照账户的注册时间获取的，按照账户的注册时间进行排序后获取到归档起点至归档终点分发区间的账户。

6.根据权利要求 1 所述的云存储中文件的归档压缩方法，其特征在于，所述步骤 S3 中，对归档起点、终点和归档服务进程特征码进行记录，当归档服务进程在提取访问次数少于文件冷热程度阀值 H 且文件小于等于归档文件大小阈值 S 的文件过程中，若出现异常退出，则回收归档服务进程执行的该任务，并且将回收的任务添加到待分发任务的列表中。

7.根据权利要求 1 所述的云存储中文件的归档压缩方法，其特征在于，所述步骤 S5 中，在压缩后的文件存储到云存储中后，在确保存储成功后将压缩后的相关信息添加到所压缩文件的原来的信息中。

8.根据权利要求 1 所述的云存储中文件的归档压缩方法，其特征在于，所述步骤 S5 中，归档服务进程定时反馈心跳信息和汇报任务完成进度。

9.根据权利要求 1 所述的云存储中文件的归档压缩方法，其特征在于，所述归档文件大小阈值 S 为 8MB，文件冷热程度阀值 H 为 100。

一种云存储中文件的归档压缩方法

技术领域

[0001] 本发明涉及云存储平台的海量文件归档压缩的技术，特别涉及一种云存储中文件的归档压缩方法。

背景技术

[0002] 云存储是在云计算概念上延伸和衍生发展出来的一个新的概念。云计算是分布式处理 (Distributed Computing)、并行处理 (Parallel Computing) 和网格计算 (Grid Computing) 的发展，是透过网络将庞大的计算处理程序自动分拆成无数个较小的子程序，再交由多部服务器所组成的庞大系统经计算分析之后将处理结果回传给用户。通过云计算技术，网络服务提供者可以在数秒之内，处理数以千万计甚至亿计的信息，达到和超级计算机同样强大的网络服务。云存储是一种服务，和云计算相似，通过集群应用、网格技术或分布式文件系统等功能，将网络中大量各种不同类型的存储设备通过应用软件集合起来协同工作，共同对外提供数据存储和业务访问功能的一整套系统。

[0003] 云计算技术在生活中的应用越来越广泛，云存储作为云计算的底层支撑，集中了云计算后端架构的难点，而云存储性能的好坏将直接影响到云计算向用户提供服务的性能，也因此影响到用户的体验。在云存储基础设施服务领域，面对存储池中的海量文件，能有效的提高云存储的存储容量、减缓存储中文件数目的增长速度的同时保证存储文件的可靠性，目前仍未找到有效可行的解决方案。目前较多的云存储服务提供商的解决方案并没有对文件进行归档压缩的处理步骤，而 Facebook 的 Haystack 云存储解决方案也只是对文件进行归档，但并没有进行压缩处理，这样虽然减缓了文件数目的增长速度，但是没有节省云存储的存储容量，而过大的单个文件出现故障时的文件迁移耗时较长并且会影响集群性能。

[0004] 而现有的云存储平台在后台存储中采用单文件多备份的方式进行存储，并将备份均衡地放置到不同的磁盘上，然而随着文件数的指数式增长，磁盘的读写性能会下降从而影响服务质量，存储空间也是固定的占用了文件的大小乘以备份数的容量，因此采取此种方式无法使得云存储的存储效益最优化。

发明内容

[0005] 本发明的目的在于克服现有技术的缺点与不足，提供一种云存储中文件的归档压缩方法，该方法根据云存储中文件访问热度调整存储方式，实现了云存储文件数目增加速度收敛和存储效益的提高。

[0006] 本发明的目的通过下述技术方案实现：一种云存储中文件的归档压缩方法，步骤如下：

[0007] S1、设置归档文件大小阈值 S 和文件冷热程度阀值 H，通过归档服务进程提出归档服务请求；

[0008] S2、根据归档服务进程的归档服务请求分发归档起点和归档终点，然后获取到处

于归档起点和归档终点分发区间的账户；

[0009] S3、在云存储对外服务闲置时段，归档服务进程执行任务：归档服务进程依次遍历处于归档起点和归档终点分发区间的账户，将各账户下访问次数少于文件冷热程度阀值 H 且文件小于等于归档文件大小阈值 S 的文件提取出来；

[0010] S4、对所提取的文件的完整性进行检验，判断所提取的各文件与原文件是否一致；

[0011] 若否，则重新提取该文件，然后针对该文件的完整性进行重新检验，直到获取到与原文件相同的文件；

[0012] 若是，则进入步骤 S5；

[0013] S5、根据文件的特性，对步骤 S4 中获取的与原文件一致的文件进行归档压缩，然后将压缩后的文件存储到云存储中，并且将云存储中对应的原文件进行删除。

[0014] 优选的，还包括以下步骤：

[0015] 设置访问时间间隔 I，当文件被访问时，判断该文件是否已归档压缩；

[0016] 若是，则查询文件的具体存储路径，然后从压缩文件中提取出目标文件并返回文件的内容；

[0017] 若否，则判断该文件当前访问时间与上次访问时间之差是否超过访问时间间隔 I，若是，则将该文件的访问次数置 1，若否，则将其访问次数加 1。

[0018] 优选的，所述访问时间间隔 I 为 15 天以上。

[0019] 优选的，还包括以下步骤：文件写入时，将其访问次数置为文件冷热程度阀值 H。

[0020] 优选的，所述步骤 S2 中，归档起点至归档终点分发区间的账户是按照账户的注册时间获取的，按照账户的注册时间进行排序后获取到归档起点至归档终点分发区间的账户。

[0021] 优选的，所述步骤 S3 中，对归档起点、终点和归档服务进程特征码进行记录，当归档服务进程在提取访问次数少于文件冷热程度阀值 H 且文件小于等于归档文件大小阈值 S 的文件过程中，若出现异常退出，则回收归档服务进程执行的该任务，并且将回收的任务添加到待分发任务的列表中。

[0022] 优选的，所述步骤 S5 中，在压缩后的文件存储到云存储中后，在确保存储成功后将压缩后的相关信息添加到所压缩文件的原来的信息中。

[0023] 优选的，所述步骤 S5 中，归档服务进程定时反馈心跳信息和汇报任务完成进度。

[0024] 优选的，所述归档文件大小阈值 S 为 8MB，文件冷热程度阀值 H 为 100。

[0025] 本发明相对于现有技术具有如下的优点及效果：

[0026] (1) 本发明方法根据文件的大小以及被访问的次数进行归档压缩，在云存储对外服务闲置时段，将小于文件大小阈值 S 以及被访问次数小于文件冷热程度阀值 H 的文件进行归档以及压缩处理，使得存储池中的文件数目会缓慢增长，相对于指数增加而言，大大地降低了文件数目的增长速度，减少了磁盘上文件的数目，提高磁盘的性能，并且节省了存储空间和存储成本，提高了存储效益。另外本发明方法在云存储对外服务闲置时段才进行归档和压缩处理，由于在存储对外服务闲置时段，计算资源使用率是很低的，而将其用于归档压缩处理则充分提高了其利用率，并节省了额外购置压缩归档处理服务器的开支。通过本发明方法对云存储中海量文件进行合理的归档压缩，能有效地提高单位存储空间里存储文

件的密度并避免磁盘上文件数过多带来的性能下降的弊端，从而进一步体现云计算的高性价比和高可靠性的优势。

[0027] (2) 本发明方法在文件被访问时，当文件当前访问时间与上次访问时间之差超过访问时间间隔 I，则将该文件的访问次数置 1，没有超过时，则将该文件的访问次数加 1，因此本发明方法将文件的访问频率考虑进去，将文件访问频率低的文件进行归档压缩。

[0028] (3) 本发明方法在新文件写入时，将其访问次数首先置为文件冷热程度阀值 H，避免新文件刚刚写入时，由于访问次数少于冷热程度阀值 H 而被误归档压缩。

附图说明

[0029] 图 1 是本发明方法流程图。

具体实施方式

[0030] 下面结合实施例及附图对本发明作进一步详细的描述，但本发明的实施方式不限于此。

[0031] 实施例

[0032] 如图 1 所示，本实施例公开一种云存储中文件的归档压缩方法，步骤如下：

[0033] S1、设置归档文件大小阈值 S 和文件冷热程度阀值 H，通过归档服务进程提出归档服务请求；其中在本实施例中归档文件大小阈值 S 为 8MB，文件冷热程度阀值 H 为 100。

[0034] S2、根据归档服务进程的归档服务请求分发归档起点和归档终点，然后获取到处于归档起点和归档终点分发区间的账户；其中，归档起点和归档终点是指按账户注册时间排序后账户区间起点和区间终点，归档起点至归档终点分发区间的账户是按照账户的注册时间获取的，按照账户的注册时间进行排序后获取到归档起点至归档终点分发区间的账户。

[0035] S3、在云存储对外服务闲置时段，归档服务进程执行任务：归档服务进程依次遍历处于归档起点和归档终点分发区间的账户，将各账户下访问次数少于文件冷热程度阀值 H 且文件小于等于归档文件大小阈值 S 的文件提取出来；其中在本步骤中，对归档起点、终点和归档服务进程特征码进行记录，当归档服务进程在提取访问次数少于文件冷热程度阀值 H 且文件小于等于归档文件大小阈值 S 的文件过程中，若出现异常退出，则回收归档服务进程执行的该任务，并且将回收的任务添加到待分发任务的列表中。

[0036] S4、对步骤 S3 所提取的文件的完整性进行检验，判断所提取的各文件与原文件是否一致；

[0037] 若否，则重新提取该文件，然后针对该文件的完整性进行重新检验，直到获取到与原文件相同的文件；

[0038] 若是，则进入步骤 S5；

[0039] S5、根据文件的特性，对步骤 S4 中获取的与原文件一致的文件分别进行归档压缩，即将这些文件中具有某些相同特性（如属于同一个账户的文件、存放时间相近的文件、大小相近的文件等特性）的一些文件存放到同一个目录下，然后对该目录进行压缩，将压缩后的文件存储到云存储中，并且将云存储中对应的原文件进行删除。本步骤中，归档服务进程定时反馈心跳信息和汇报任务完成进度，其中心跳信息就是归档服务进程进行其运行

状态是否正常的一种汇报的信息。在压缩后的文件存储到云存储中后，在确保存储成功后将压缩后的相关信息添加到所压缩文件的原来的信息中。其中，压缩后的相关信息是指压缩前的文件现在是压缩文件的第几个文件的位置信息以及该压缩文件的具体存储路径。压缩文件原来的信息是指在压缩前云存储中记录的该文件的存储路径、文件大小、文件名称和文件的校验和等文件信息。

[0040] 本实施例方法还包括以下步骤：

[0041] 设置访问时间间隔 I，当文件被访问时，判断该文件是否已归档压缩；

[0042] 若是，则查询文件的具体存储路径，即找到具体哪台机器上哪个磁盘上的哪个目录下的哪个压缩文件里面的第几个文件，然后从压缩文件中提取出目标文件并返回文件的内容；

[0043] 若否，则判断该文件当前访问时间与上次访问时间之差是否超过访问时间间隔 I，若是，则将该文件的访问次数置 1，若否，则将其访问次数加 1。

[0044] 在本实施例中访问时间间隔 I 为 15 天，当然也可以为 15 天以上或者其他合适的天数。

[0045] 在本实施例中方法中文件写入时，将其访问次数置为文件冷热程度阀值 H。避免新文件刚刚写入时，由于访问次数少于冷热程度阀值 H 而被误归档压缩。待该新文件当前访问时间与上次访问时间之差超过访问时间间隔超过 I 时，其访问次数被置为 1，此时由于其被访问的频率下降，而有可能被归档压缩。因此本实施例方法将文件的访问频率考虑进去，将文件访问频率低的文件进行归档压缩。

[0046] 上述实施例为本发明较佳的实施方式，但本发明的实施方式并不受上述实施例的限制，其他的任何未背离本发明的精神实质与原理下所作的改变、修饰、替代、组合、简化，均应为等效的置换方式，都包含在本发明的保护范围之内。

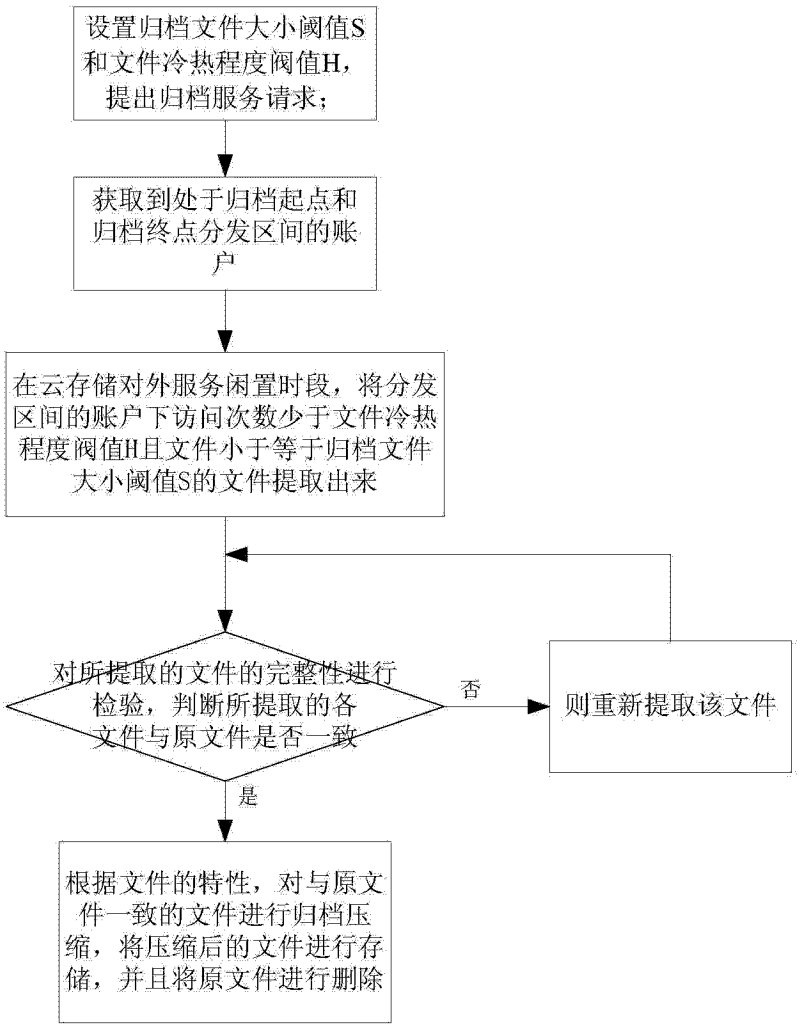

设置归档文件大小阈值S和文件冷热程度阀值H，提出归档服务请求；
获取到处于归档起点和归档终点分发区间的账户
在云存储对外服务闲置时段，将分发区间的账户下访问次数少于文件冷热程度阀值H且文件小于等于归档文件大小阈值S的文件提取出来
对所提取的文件的完整性进行检验，判断所提取的各文件与原文件是否一致
否
则重新提取该文件
是
根据文件的特性，对与原文件一致的文件进行归档压缩，将压缩后的文件进行存储，并且将原文件进行删除

图 1