(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 103473258 A
(43)申请公布日 2013.12.25

(21)申请号 201310232354.4

(22)申请日 2013.06.01

(71)申请人 西安邮电大学
地址 710061 陕西省西安市长安南路 563 号

(72)发明人 陈莉君 康华 贾威威 王博

(51)Int.Cl.
G06F 17/30(2006.01)
G06F 11/14(2006.01)

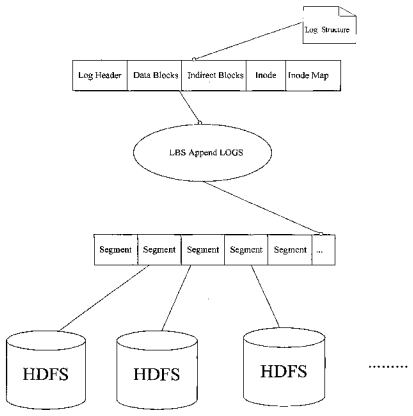
权利要求书1页 说明书4页 附图1页

(54)发明名称

云存储文件系统

(57)摘要

本发明公开了一种云存储文件系统，在Hadoop分布式文件系统基础之上实现日志结构块存储系统；所述Hadoop分布式文件系统，用于提供分布式的存储介质；所述日志结构块存储系统包括快照单元，克隆单元，块压缩单元，缓存单元等等。本发明的优点在于通过封装HDFS(Hadoop Distributed File-System)接口实现HLBS(Hadoop Log-Structured Block-Storage System)的接口，在此基础上实现LBS(Log-Structured Block-Storage System)，最终实现创新型Hadoop分布式日志结构块存储系统，从而实现对数据的随机读写以及I/O效率的提高，同时包含快照，克隆等功能。



CN 103473258 A

1. 一种云存储文件系统，其特征在于，包括：Hadoop 分布式文件系统和日志结构块存储系统；所述 Hadoop 分布式文件系统，用于提供分布式的存储介质；所述日志结构块存储系统包括随机读写单元、快照单元、克隆单元、块压缩单元和缓存单元。

2. 根据权利要求 1 所述的一种云存储文件系统，其特征在于，所述 Hadoop 分布式文件系统包括：接收分布式文件系统客户端发送的访问请求；根据所述访问请求确定所述分布式文件系统客户端所要访问的存储单元以及所述存储单元对应的虚拟资源池状态，所述存储单元与其对应的虚拟资源池存储有相同的数据，所述虚拟资源池为根据具有相同活动周期的分布式文件系统客户端聚类生成；在确定所述存储单元对应的虚拟资源池处于上线状态时，向发起访问请求的分布式文件系统客户端返回包括客户端标识的访问响应消息，所述客户端标识所标识的分布式文件系统客户端为所述虚拟资源池中的一个，且根据所述访问请求确定。

云存储文件系统

技术领域

[0001] 本发明涉及一种数据存储技术，特别涉及一种云存储文件系统。

背景技术

[0002] HDFS 是一个高度容错性的系统，适合部署在廉价的机器上。HDFS 能提供高吞吐量的数据访问，非常适合大规模数据集上的应用。HDFS 一般部署在集群环境中，而且这个集群环境是一个主从 (master/slave) 系统。在这个系统中有一个命名节点 (Namenode) 和若干个数据节点 (Datanode)，命名节点上存储着所有元数据 (Meta-data)，而数据节点上存储着所有用户数据，这些数据被组织成数据块的形式放在数据节点上，每个数据块默认存放在三个数据节点（可配置）上，用户的请求（读写等）都是通过命名节点，进而操作数据节点。HDFS 一次写入不能更改，只可多次读取。一旦创建了一个 HDFS 文件，并且写入了数据，关闭之后就不能再修改这些数据了。这种方式简化了数据一致性，同时也使高数据吞吐量变为可能。但是，这种方式也限制了用户对数据的操作，这是 HDFS 的一个鲜明的缺点。

[0003] 日志结构文件系统 (Log-Structured File System) 最早由 John K. Ousterhout 和 Fred Douglis 在 1988 年提出。这种设计是为了提高写数据吞吐量，所有对数据和元数据的更新都是以日志的形式追加，形成一个线性的数据结构。日志结构文件系统 (LFS) 会不断的追加日志结构，日志中元数据的数据量有可能大于可用数据，每次更新都会产生元数据，那么存储空间就成为我们关注的核心问题，这么多的冗余数据应该如何处理。这就是日志结构文件系统的缺点。

[0004] 虽然，中国专利 CN201010624684.4 公开了一种分布式文件系统的数据存储处理方法，其特征在于，包括 ：接收分布式文件系统客户端发送的访问请求 ；根据所述访问请求确定所述分布式文件系统客户端所要访问的存储单元以及所述存储单元对应的虚拟资源池状态，所述存储单元与其对应的虚拟资源池存储有相同的数据，所述虚拟资源池为根据具有相同活动周期的分布式文件系统客户端聚类生成 ；在确定所述存储单元对应的虚拟资源池处于上线状态时，向发起访问请求的分布式文件系统客户端返回包括客户端标识的访问响应消息，所述客户端标识所标识的分布式文件系统客户端为所述虚拟资源池中的一个，且根据所述访问请求确定。上述技术方案能够提高分布式文件系统的系统稳定性，但并没能进一步解决冗余数据问题。类似的专利技术还有很多，但都程度不同地存在着稳定性和实用性不佳的问题，尚有待于进一步改进完善。

发明内容

[0005] 本发明的目的在于提供一种云存储文件系统，可以实现对数据的随机读写以及 I/O 效率的提高，从而解决上述问题。

[0006] 为实现上述发明目的，本发明的技术方案是 ：一种云存储文件系统，包括 ：Hadoop 分布式文件系统和日志结构块存储系统 ；所述 Hadoop 分布式文件系统，用于提供分布式的存储介质 ；所述日志结构块存储系统是基于日志结构文件系统理念而实现随机读写，它还

包括快照单元，克隆单元，块压缩单元，缓存单元。

[0007] 作为本发明的优选实施例，所述 Hadoop 分布式文件系统包括：接收分布式文件系统客户端发送的访问请求；根据所述访问请求确定所述分布式文件系统客户端所要访问的存储单元以及所述存储单元对应的虚拟资源池状态，所述存储单元与其对应的虚拟资源池存储有相同的数据，所述虚拟资源池为根据具有相同活动周期的分布式文件系统客户端聚类生成；在确定所述存储单元对应的虚拟资源池处于上线状态时，向发起访问请求的分布式文件系统客户端返回包括客户端标识的访问响应消息，所述客户端标识所标识的分布式文件系统客户端为所述虚拟资源池中的一个，且根据所述访问请求确定。

[0008] 在本发明中，HLBS 的磁盘数据格式与一般文件系统无多大差异，都是借助于 data block、indirect block、inode 等结构。所不同之处在于 LBS 会将磁盘（这里是 HDFS 的存储池）分割成有序的 segment 进行管理，当前活跃的 segment 只有一个（也就是日志的逻辑尾的 segment）。这些 segment 逻辑上头尾相连组成线性 logs，任何对文件的更新（data block、indirect block、inode 等等）都会以追加方式写入一个新的 log——显然这么做的好处是保证了磁头的顺序移动，提高了吞吐量；而带来的麻烦是需要回收前期写入的旧数据（修改过的），否则磁盘迟早会写满。综上所述我们设计的基本思路是——利用 HDFS 为我们提供可靠的、分布式的存储介质；然后在其上实现 LBS。

[0009] 其中 log 是我们数据持久化的一个基本写入单位，对于写透需求来说，实际上每次写入动作都会产生一个新的 log，而每次的 log 大小不尽相同。log 的内容显然必须包含被写入的数据块，还需要包含对应的元数据（索引块等）信息，以及元数据的元信息（inode），这样才能完成对数据的索引。任何文件或者目录的修改，LBS 都需要向 log 中写入如上几部分信息，而且要求严格“按照顺序写入（in-order semantics）”——其目的是为了崩溃时能尽可能恢复数据一致性。

[0010] 读取文件最新数据时需要通过找到最新的 inode map 位置，再进而找到所需文件对应 inode，再进而找到文件逻辑地址对应的数据块的物理地址（段号 +offset），再进而读取数据。最新的 Inode map 位置理应记录在 checkpoint 文件中，HDFS 初始化加载时读入；如果运行中则该 inode map 驻留于内存数据结构中。文件块大小是可变的（可配置），比如 8k。对于不足一个块的修改，一定会伴随先读出完整块再修改，再追加这一过程。

[0011] 具体来说，LBS 是基于 LFS 的理念，但是又不同于 LFS。LBS 简化了 LFS，通过 LFS 的设计理念设计并实现了块级别的日志存储系统，同时在 LBS 之上实现了快照（线性快照和树形快照），克隆，块压缩，缓存等技术。同时，对 HDFS 也进行了改进，实现了副本迁移，在无网络环境下，可以实现本地启动虚拟机等功能。

[0012] 采用了上述技术方案，本发明的有益效果为：通过封装 HDFS 接口，实现 HLBS 的接口，在此基础上实现 LBS，最终实现创新型 Hadoop 分布式日志结构块存储系统，HLBS 吸取了 HDFS 和 LBS 的各自优点，同时弥补了各自的缺点，从而实现对数据的随机读写以及 I/O 效率的提高，同时包含快照，克隆等功能。

附图说明

[0013] 图 1 为本发明 HLBS 实现原理图。

具体实施方式

[0014] 下面结合实施例对本发明进一步说明。

[0015] 实施例：一种云存储文件系统，包括：Hadoop 分布式文件系统和日志结构块存储系统；所述 Hadoop 分布式文件系统，用于提供分布式的存储介质；所述日志结构块存储系统是基于日志结构文件系统而实现随机读写，还包括快照单元、克隆单元、块压缩单元和缓存单元。其中：所述 Hadoop 分布式文件系统包括：接收分布式文件系统客户端发送的访问请求；根据所述访问请求确定所述分布式文件系统客户端所要访问的存储单元以及所述存储单元对应的虚拟资源池状态，所述存储单元与其对应的虚拟资源池存储有相同的数据，所述虚拟资源池为根据具有相同活动周期的分布式文件系统客户端聚类生成；在确定所述存储单元对应的虚拟资源池处于上线状态时，向发起访问请求的分布式文件系统客户端返回包括客户端标识的访问响应消息，所述客户端标识所标识的分布式文件系统客户端为所述虚拟资源池中的一个，且根据所述访问请求确定。HLBS 是一个在 HDFS 文件系统之上实现的 LBS 系统。但要注意它并非一个实现完整 POSIX 语义的文件系统（支持目录操作、link 等），目前实现了单一文件的基本管理 (open, write, read, close) 的系统，所以把它称为存储系统 (block-level) 可能更加合适，HLBS 已经支持了很多高级存储管理技术，比如，快照，克隆，块压缩，缓存等，同时，HLBS 也已经支持了目前业内比较有名的系统，如 XEN，QEMU/KVM，Libvirt，Openstack 等。

[0016] HLBS 快照技术的主要作用是能够进行在线数据备份与恢复。当存储设备发生应用故障或者文件损坏时可以进行快速的数据恢复，将数据恢复到某个可用的时间点的状态。快照的另一个作用是为存储用户提供另外一个数据访问通道，当原数据进行在线应用处理时，用户可以访问快照数据，还可以利用快照进行测试等工作。HLFS 快照技术包括线性快照和树形快照，树形快照应用场景更为广泛。

[0017] HLBS 克隆技术主要作用是在一个 HDFS 之上的系统盘镜像，可以作为无数新系统的 base 系统，从而提高系统新系统生产速度和解决存储空间。同时，为了减少本网络传输压力，提高系统响应速度，可考虑利用本地文件系统作为 Base 数据宿主：比如一些场景中——我们可以将标准的镜像或某系统通用软件做到工具盘中，并置于到本地 HLBS 上，即 local 方式挂载的 HLBS 系统上，然后再在集群的 HDFS 上做一个新 HLBS 系统，并将其 base 系统放到本地的上述 HLBS 系统上。从而只有变化的增量数据需要途径网络 I/O，这样很大程度上会提高系统性能。HLBS 块压缩技术主要作用是通过压缩算法来重新组织数据，使存储空间得到最大利用。HLBS 缓存技术主要作用是为了提高 I/O 效率。HLBS 支持 XEN 虚拟机，在 XEN 虚拟机中可以创建 HLBS 卷，从而使 XEN 虚拟机具备了 HLBS 的所有优良特性。HLBS 还支持 QEMU/KVM 虚拟机，在 QEMU/KVM 虚拟机中可以创建 HLBS 卷，进而对 HLBS 卷应用快照，克隆，块压缩，缓存等技术。HLBS 支持 XEN，QEMU/KVM 虚拟机之后，使虚拟机的性能得到了极大的提高。Libvirt 提供了一套标准的虚拟化接口，HLBS 支持 Libvirt，使用者可以通过 Libvirt 创建 HLBS 卷，进而应用 HLBS 所提供的一系列功能。Openstack 是一个云基础软件，目前在业界非常有影响力，HLBS 支持 Openstack，可以通过 Openstack 创建 HLBS 卷，进而把 HLBS 的所有特性集成到 Opesntack 项目中，使 Openstack 更加强大。

[0018] 综上所述，HLBS 在 HDFS 之上实现 LBS，从而达到随机读写，存储空间扩展等特性。HLBS 还支持快照，克隆，块压缩等技术。目前，HLBS 已经支持 XEN，QEMU/KVM，Libvirt，

Openstack 等著名项目。HLBS 的应用场景和范围越来越广泛。

[0019] 本发明不局限于上述具体的实施方式，本领域的普通技术人员从上述构思出发，不经过创造性的劳动，所作出的种种变换，均落在本发明的保护范围之内。

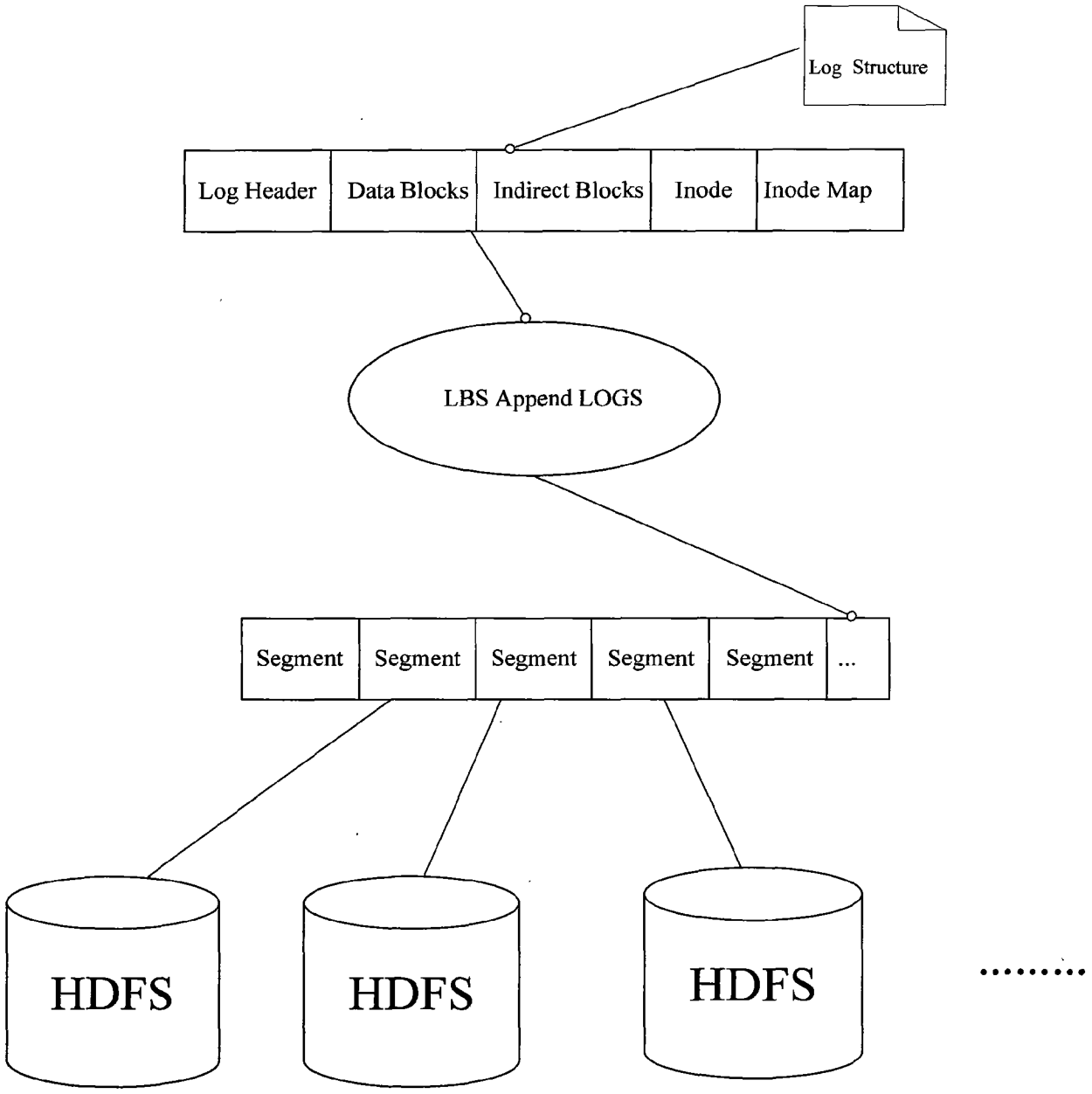

Log Structure
Log Header
Data Blocks
Indirect Blocks
Inode
Inode Map
LBS Append LOGS
Segment
Segment
Segment
Segment
Segment
...
HDFS
HDFS
HDFS
.....

图 1