

面向云存储的归档方法

范中磊

(长安大学 信息工程学院, 陕西 西安 710064)

摘 要: 针对目前云存储系统主要提供文件备份的现状, 提出了一种面向云存储的归档方法, 该方法将不活跃的文件数据以归档的形式存放在云端存储系统, 并在客户端保留其存根信息以供读取时使用, 从而不仅大大节约了客户端存储空间, 改善了客户端性能, 而且实现了对用户透明的归档数据访问。

关键词: 云存储; 文件备份; 文件归档; 按需存储; 文件系统

中图分类号: TP333

文献标识码: A

文章编号: 1000-7180(2011)07-0136-03

Cloud Storage-Oriented Archiving Approach

FAN Zhong-lei

(School of Information Engineering, Chang'an University, Xi'an 710064, China)

Abstract: Aiming at the situation that current cloud storage systems mainly focus on file backup, this paper presents a cloud storage-oriented archiving approach. By storing inactive file data in cloud storage systems in archival form and leaving its stub information on clients for the use of read, this approach can not only save storage space on clients and improve the performance, but also implement transparent user access to archived data.

Key words: cloud storage; file backup; file archiving; storage-on-demand; file system

1 引言

随着信息技术,特别是网络技术的迅速发展,数据量正以前所未有的速度急剧增长. 如何对快速增长的海量数据进行安全、高效、低成本的存储、管理和使用已成为目前存储系统面临的巨大挑战。

云存储作为一种新形态存储系统,以其在大容量、易管理、高可扩展和按需服务等方面的优势而受到学术界和产业界的重视,被普遍认为是应对未来存储系统所面临挑战的一种有效方法^[1-5]. 云存储是一种基于 Internet 的大规模“集中”式存储系统,其存储系统是由地理上广泛分布的各种不同类型的存储设备构成,由专门的存储服务供应商(Storage Service Provider, SSP)管理. 云存储以数据存储服务的形式提供给用户,简化了用户的部署和使用成本,使能按需的、绿色的存储。

目前云存储系统大都以文件备份的形式提供数据存储服务. 尽管后端的云存储系统提供了巨大的

存储空间,但随着文件和所需存储数据的增多,前端客户端的存储空间会越来越小,性能也会随之下降. 而另一方面,客户端上又存放着大量长期不被访问的文件,占据着存储空间. 据全球网络存储工业协会(Storage Network Industry Association, SNIA)数据估计,存储系统上的数据中有 68% 的数据在 3 个月或更长的时间里处于“冬眠”状态,有高达 80% 的数据在一年以后不再被访问或修改^[6].

针对这种情况,文中提出了一种面向云存储的归档方法,该方法将不活跃的文件数据以归档的形式存放在云存储系统,并在客户端保留其存根信息以供读取时使用,从而大大节约了客户端存储空间,改善了客户端性能,而且实现了对用户透明的归档数据访问。

2 设计思想

备份用于创建数据副本,当原始数据丢失或者损坏且无法修复时,副本就可用于恢复原始数据. 备

份通常用于保护正在使用的数据. 备份只是复制数据, 原始文件仍然保留在原地, 并不能释放存储空间. 即备份使用的存储空间是原来的两倍, 除非在备份以后手动删除文件, 这又增加了数据管理的开销.

归档将数据从其主存储系统迁移到归档存储系统中, 用于长期保留并供将来使用. 归档通常用于保护停止改变或较少使用的数据. 归档是迁移数据, 文件在原地已经不存在了, 因此其并不增加使用的存储空间.

目前, 在线文件备份是云存储的主要应用之一. 通过备份应用软件将数据副本存放到地理上广泛分布的云存储系统中, 可以有效地实现数据保护和灾难恢复. 尽管云存储提供了海量的存储空间, 但是随着时间的推移和文件的增多, 作为云存储前端的客户端的存储空间会越来越小, 性能也会越来越低. 另一方面, 客户端文件系统中存在着大量的不活跃数据(即较少使用的数据), 占据着宝贵的存储空间.

采用传统的归档方法, 将不活跃的数据直接从客户端删除, 而只留一份数据在云存储系统, 不能实现对用户透明, 无法满足云存储的用户需求. 因此, 需要对传统的归档方法进行扩展, 寻求一种适用于云存储的归档方法. 通过将文件数据从客户端删除后, 保留一个信息存根的方法, 可以方便的实现透明的归档数据访问.

3 软件结构与工作流程

图1给出了面向云存储的归档方法的软件结构与工作流程. 图1(a)描述了文件归档的流程:

(1) 文件备份模块从本地文件系统读取需要备份的文件;

(2) 文件备份模块通过云存储接口在云存储系统中创建备份文件副本;

(3) 文件备份模块写数据库表, 记录文件存储状态为已备份文件, 即本地和云存储系统都存放有该文件的数据. 当需备份的文件发生变化时, 文件备份模块会重复上述步骤, 建立新的文件副本;

(4) 文件归档模块定期的扫描本地文件系统, 根据用户指定的归档策略, 将需要归档的文件数据从本地删除, 而只保留该文件的信息存根, 从而释放出本地存储空间;

(5) 文件归档模块改写数据库表, 记录文件存储状态为已归档文件, 即文件数据只在云存储系统中存在, 而本地保留信息存根.

图1(b)描述了归档文件的读取流程:

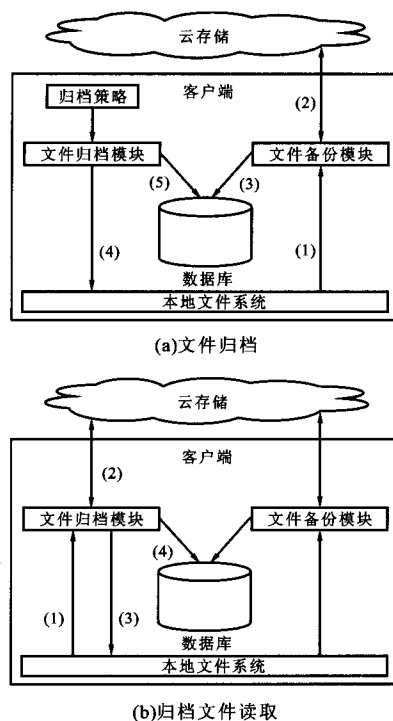


图1 面向云存储的归档方法的软件结构与流程

(1) 当用户访问已归档文件时, 发送文件读取请求到文件归档模块;

(2) 文件归档模块通过云存储接口与云存储系统进行交互, 获取归档文件信息;

(3) 文件归档模块将所请求文件数据写入本地文件系统, 以供用户访问;

(4) 文件归档模块改写数据库表, 记录文件存储状态为已备份文件, 即本地和云存储系统都存放有该文件的数据. 上述文件数据读取过程对用户是透明的, 即用户感觉就像使用本地文件系统中的文件一样, 并不知道该文件是从云存储中读取的, 只是有时会觉得稍有延迟.

4 系统测试

采用上述方法, 在 Linux 环境开发实现了一个原型系统. 在该系统中, 采用一台服务器作为云端的存储服务器, 另外一台服务器作为客户端. 测试的基本过程包括:

(1) 运行预处理脚本. 主要完成两项工作: ①在客户端上创建一个 40MB 的磁盘分区; ②建立用于测试的文件样本集(包括 40MB、20MB、10MB、2MB 四种类型, 分别用于模拟实际的不同工作环境);

(2) 运行测试脚本, 拷贝文件到上述 40MB 的磁盘分区. 该测试脚本允许选择文件样本集的类型和

所需拷贝样本集的数量. 例如, 选择 2MB 的文件样本集, 将文件样本集的数量选为 200, 这样就会有 400MB 的文件数据拷贝到 40MB 的磁盘分区; 当系统检测到磁盘分区的已用空间超过某一阈值时, 便根据归档策略选择文件进行归档操作, 将需要进行归档的文件拷贝到云端的存储服务器, 然后释放出客户端磁盘分区的存储空间, 以容纳更多的文件数据;

(3) 运行文件读取脚本, 读取被归档的文件, 看可否正常读取. 对于已经归档的文件, 系统会自动从云端的服务器拷贝本地客户端服务器, 然后执行客户端的读文件命令, 客户端就像使用本地文件一样使用文件, 并不会知道文件是从远端的服务器读取的, 只是有时会觉得稍微有点延迟.

表 1 给出了部分测试用例, 测试结果表明: 系统不仅可以正确、可靠的运行, 而且可以为客户端节约 10 倍以上的存储空间.

表 1 测试用例(磁盘分区大小为 40MB)

文件样本 集类型(MB)	文件样本 集数量(个)	文件样本 数据大小(MB)	节约客户端 存储空间(倍)
40	10	400	10
20	20	400	10
10	50	500	12.5
2	200	400	10
2	300	600	15

5 结束语

文中所述方法在原型系统得以实现和验证, 可以有效地解决云存储中备份客户端由于数据量的增长而导致的存储空间不足和性能下降的问题. 但是还需要进一步的研究和完善, 例如: 如何支持客户端集群的归档需求、如何提供更加有效的归档策略等.

参考文献:

- [1] 巴力立. 企业数据中心改革: 从云计算到云存储 [EB/OL]. (2009-08-19)[2010-03-13]. <http://www.chinacloud.cn/show.aspx?id=2451&cid=30>.
- [2] 曾龙海, 张博锋, 张丽华, 等. 基于云计算平台的虚拟集群构建技术研究[J]. 微电子学与计算机, 2010, 27(8): 31-35, 40.
- [3] 王鹤群. 云存储的应用 [J]. 记录媒体技术, 2008 (5): 60-64, 51.
- [4] 胡金鹏, 张延园, 姜楠, 等. 远程数据复制系统的日志保护机制研究[J]. 微电子学与计算机, 2010, 27(11): 117-120.
- [5] 李煜民, 章才能, 谢杰. 云计算环境下的数据存储[J]. 电脑知识与技术, 2010, 6(5): 1032-1034.
- [6] Enterprise Strategy Group (ESG). Hitachi Content Archive Platform [EB/OL]. (2006-06-01)[2010-03-13]. <http://www.hds.com/assets/pdf/esg-report-on-the-hitachi-content-archive-platform.pdf>.

作者简介:

范中磊 男, (1970-) 博士, 教授. 研究方向为分布式存储与计算.

(上接第 135 页)

- [3] 吴卫国, 陈辉堂, 王月娟. 移动机器人的全局轨迹跟踪控制[J]. 自动化学报, 2001, 27(3): 326-331.
- [4] 叶涛, 侯增广, 谭民, 等. 移动机器人的滑模轨迹跟踪控制[J]. 高技术通讯, 2004(11): 71-74.
- [5] Fukao T, Nakagawa H, Adachi N. Adaptive tracking control of a nonholonomic mobile robot [J]. IEEE Transactions on Robotics and Automation, 2000, 16 (5): 609-615.
- [6] Kim M S, Shin K H, Hong S G, et al. Designing a robust adaptive dynamic controller for nonholonomic mobile robots under modeling uncertainty and disturbances [J]. Mechatronics, 2003, 13: 507-519.
- [7] 闫茂德, 贺显曜, 吴青云. 移动机器人全局轨迹跟踪的自适应滑模控制[J]. 微电子学与计算机, 2006, 23(4): 97-100.

- [8] 程俊林, 赵东风. 轮式移动机器人滑模轨迹跟踪控制 [J]. 云南大学学报: 自然科学版, 2007, 29(S1): 64-67.
- [9] 李传峰, 王朝立. 一类不确定非完整运动学系统的鲁棒镇定[J]. 北京航空航天大学学报, 2007, 33(4): 427-431.
- [10] Kristic M, Kanellakopoulos, Kokotovic P V. Nonlinear and adaptive control design [M]. New York: Wiley, 1995.

作者简介:

张鑫 男, (1974-), 博士研究生, 研究方向为移动机器人自主导航与轨迹跟踪控制.

闫茂德 男, (1985-), 博士研究生, 研究方向为非线性系统控制和机器人学.