

基于异步数据云的云备份平台^①

姜学峰¹, 许霄波², 吴放¹, 董惠良¹

¹(浙江中烟工业有限责任公司, 杭州 310009)

²(浙江创联信息技术股份有限公司, 杭州 310012)

摘要: 随着信息化技术的发展, 信息系统安全在企业管理中显得越来越重要。在分析了云存储技术架构的基础上, 提出了一种新的基于异步数据云的系统备份平台, 将逻辑数据云与物理数据云接入网络, 提供稳定、高效的数据备份功能, 并针对不同备份对象制定备份策略, 从而达到备份资源优化的目的。在浙江中烟备份系统的应用实施, 充分说明该平台是一种适用于跨地域、大数据量的数据备份方案。

关键词: 云存储; 云备份; 异步数据云

Backup Platform Based on Asynchronous Data Clouds

JIANG Xue-Feng¹, XU Xiao-Bo², WU Fang¹, DONG Hui-Liang¹

¹(China Tobacco Zhejiang Industrial Co. Ltd, Hangzhou 310009, China)

²(Zhejiang Crea-Uniong Information technology Co. Ltd, Hangzhou 310009, China)

Abstract: With the development of the information technology, the security of information system becomes more and more important in enterprise management. In this paper, a new backup platform based on asynchronous data cloud is proposed to connect application system with logical data cloud and physical data cloud and they can back up the system data stably and efficiently. And in order to optimize the backup resources, different strategies were applied to different objects in the platform. Implement in backup system of China Tobacco Zhejiang Industrial CO.LTD shows that the platform is suit to regional and large enterprise which has great deal of data.

Key words: cloud storage; cloud backup; asynchronous data cloud

计算机与网络的发展, 推动了整个社会的数字化与信息化, 但信息系统文件与数据的安全性问题也尤为突出。面对信息化安全威胁, 如何快速而有效的对系统文件进行备份和恢复, 在计算机信息安全领域已成为研究热点^[1]。传统的备份技术, 如快照技术、磁带备份、RAID 等, 很难应付大规模的灾难, NAS、SAN 和 iSCSI 等远距离存储技术的效率和稳定性相对较低^[2,3]。所以, 现有备份技术需要进行整合和改进, 以适应企业多方面的需求。为了对系统进行备份和恢复, 必须在系统中引入冗余, 但过多的冗余会降低系统工作效率, 因此在实际应用中必须有一种基础性平台, 可以提供合适的数据冗余度和方便的服务弹性, “云存储”便是一个优秀的候选平台^[4]。云存储模型是以云计算为基础架构, 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一个系统。

本文在分析现有云存储平台技术的基础上, 针对企业的发展现状和对安全的特殊需求, 提出了一种基于异步数据云的云备份平台方案, 并在浙江中烟备份系统中予以实施。

1 云备份平台研究

1.1 云存储技术

云存储技术是随着云计算技术的发展而发展起来的。它通过集群应用、网格技术或分布式文件系统、虚拟存储等技术, 将网络中大量各种不同类型的存储

① 收稿时间:2011-11-25;收到修改稿时间:2012-01-06

设备通过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^[5]。它的基本思想是通过技术手段，将很多的资源统一起来管理，根据需要向不同的用户提供服务，从而得到资源共享及配制优化的目的。云存储系统是个复杂的系统，它由网络设备、存储设备、服务器、应用软件、公用访问接口、接入网和客户端程序等多个部分组成^[6]。

云存储技术将存储作为一种服务来提供，它与传统的存储技术相比最大的优势是存储容量大，并可方便地动态扩展和缩小容量。这避免了由于存储资源不够而无法提供优质服务或者由于资源空置造成的浪费，降低了成本。另外将数据存储进行专业的处理，简化了数据管理的工作，使企业或用户可以更专注于业务。同时借助云存储服务，数据可以存放在异地，这比将所有数据存放在同一地点能更好地抵抗灾难的发生^[7]。

1.2 云备份平台

云备份平台借鉴了云存储的原理，在网络中部署备份服务器，并将网络设备进行整合和动态分配，实现统一调度、统一备份、智能化管理、高效率恢复等数据中心级备份系统要求。与传统的云备份系统相比，基于异步数据云的云备份平台不仅提供了数据备份服务、网络存储空间，还提供了基于虚拟化存储技术和云存储管理平台的资源动态扩展能力、资源池共享架构以及备份数据异步传输机制，提高了备份的速度和有效性。该平台的核心是由逻辑数据云和物理数据云组成的异步数据云。逻辑数据云由一系列分布在不同地区的逻辑存储设备（如虚拟磁带库）组成，可以提供快速、稳定的大规模数据存储；物理数据云由多个物理存储设备（如磁带机）组成，可以提供大容量、长期稳定的存储空间。逻辑数据云和物理数据云由云备份服务器控制，根据备份策略的分配，提供并行的数据存储服务。云备份平台逻辑功能结构如图 1 所示，物理结构及主要技术将在第 2 节中详述。



图 1 云备份平台的体系结构

1) 接口层是平台的最上层，与现有应用系统和管理系统连接，提供在线备份和恢复服务，包括数据备份与应急恢复、备份状态查询、远程文件共享、集中存储、空间分配、数据归档等，并向管理系统传送操作状态和结果，接口可以采用 HTTP/HTTPS、FTP、Web Service 等协议。

2) 网络交换层是云备份平台的核心，可以是 FC 光纤通道 SAN、NAS 和 iSCSI 等 IP 存储设备，也可以是 SCSI 或 SAS 等 DAS 存储设备。这些存储设备数量庞大且分布在多个不同地域，彼此之间通过广域网、互联网或者 FC 光纤通道网络连接在一起，提供存储、网络/安全的基本设备，包括硬件和相关基础软件。网络交换层基于云备份的管理模式，主要负责整个备份系统的管理，包括对逻辑数据云中的设备进行管理和调度、备份策略的制订、备份数据库的保存等。

3) 逻辑数据云层由连入网络的若干虚拟磁带库（VTL）设备组成，主要负责系统文件和数据的快速存储、存储空间优化和已备份数据/文件的管理等，通过部署的备份软件统一调度，把目前核心业务系统的所有文件和数据均存备份到虚拟磁带库上面，实现数据中心核心生产数据的近线数据备份，同时虚拟磁带库上配置的在线重复删除技术可以实现对现有数据的快速备份/恢复功能。

4) 物理数据云层由连入网络的若干大容量物理磁带库组成，负责把虚拟磁带库备份后的数据归档到物理磁带库上面，并可以实现磁带的出库及长期保存，另外实现存储设备的管理和监控、数据统计等功能。

2 云备份平台在浙江中烟的实践研究

本节将以浙江中烟备份系统为例，详细介绍云备份平台的设计与实施。

2.1 企业概况

浙江中烟目前完成了全省卷烟工业企业整合，经济效益已进入全国卷烟工业企业前列，实现了规模和效益的同步增长。随着浙江中烟 ERP 系统（SAP）的逐步建设完善，OA 系统、PDM 系统等各应用子系统的成功投入运行，企业已进入快速发展期。信息化步伐快速提升，每个应用都在大量的处理自己的核心业务数据。为保证企业的核心业务正常运行，需要一套行之有效的系统数据存储和备份方案。

2.2 备份系统现状和需求

浙江中烟现有主机、存储系统设备包括：IBM 光纤交换机、两台 IBM DS8300 磁盘存储、两台 IBM P595、若干 P550 和 PC 服务器，现有数据实际存储容量近 20TB，主要规划为：ERP 系统全部数据，OA 部分数据，以及即将建设的 CRM、SCM、WMS 数字化仓库、批次管理系统等数据。当前状态下，一次性全备份所需要时间近 22 小时。

根据对现有设备以及基础环境的建设分析可知在备份系统的建设方面存在诸多不足：

- 1) 所有应用程序缺乏统一的备份规划，没有实现集中备份管理机制。
- 2) 目前所做的数据库备份均为本地备份，如果出现硬件损坏时，无法保存备份数据，更加无法在出现故障时候及时快速地恢复数据。
- 3) 缺乏统一的存储网络管理工具，无法有效管理集中整个存储系统。
- 4) 单次备份时间过长，备份管理系统和备份策略需改进。

2.3 系统拓扑结构

在前述云备份平台的理论研究基础之上，针对浙江中烟的数据量大、地区分布广等实际情况和对系统稳定性、灵活性等需求，云备份平台中心网络采用传输速率高、便于扩展和管理的 SAN 方式；备份服务器采用 HP DL380G7，备份系统为 Symantec NBU 7.0，并在其基础上进行二次开发；逻辑数据云由若干台 IBM TS7650G 虚拟磁带库组成，由备份管理软件进行管理和优化，同时配备 IBM 重复数据删除方案以优化备份数据；物理数据云则由若干台 IBM TS3200 磁带库组成。所有设备采用以交换机为中心的混合式 SAN 来连接，其中 SAN 服务器和备份 SAN 服务器与 SAN 交换机直接连接，RAID 磁盘阵列可使用 FC-AL 或 SATA 与交换机连接，虚拟磁带库、磁带机等存储设备可采用 FC 或 iSCSI 连接到交换机上。这种混合式 SAN 结构可以提供失效隔离、多级服务、多播传输以及可管理性等优良的网络性能。系统备份时的数据传输采用异步传输模式，备份时不影响系统的正常运行，并且节约网络资源，由于逻辑数据云与物理数据云的更新频率不同，其中的备份数据也处于异步状态。系统拓扑结构如图 2 所示。

根据前期调研情况，结合浙江中烟的实际业务需

求，通过对国内 IT 市场主流备份软件进行对比分析，系统主备份软件采用 Symantec NBU 7.0，该系统在兼容性、功能上能满足基本的管理需求，并能方便的支撑下一步容灾系统的建设。Symantec NBU 运行在主备份服务器上，对逻辑数据云中的存储设备和物理数据云中的存储设备进行方便的管理，同时有效的实施既定备份策略，完成预定的备份任务。

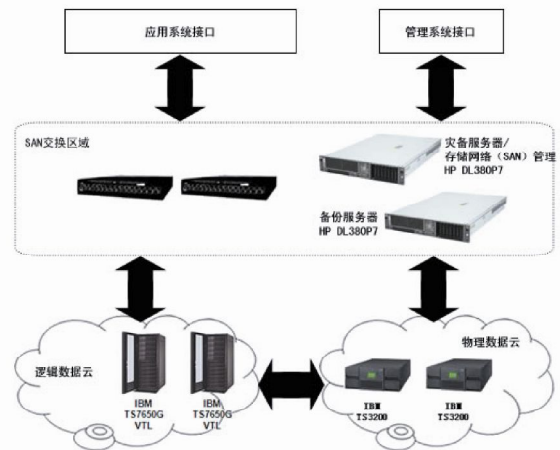


图 2 系统拓扑结构图

2.4 系统关键技术

在实施过程中，由于 Symantec NBU 只提供了基本的系统备份所需的功能，并不能完全满足浙江中烟的所有应用需求，因此进行了部分功能的二次开发，主要二次开发的功能如图 3 所示。

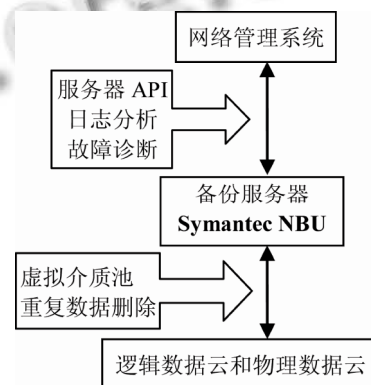


图 3 二次开发功能图

1) 备份服务器 API

为方便用户管理备份服务器及数据、监控服务器状态，在备份服务器上设置远程调用的 API，管理员或其他人员可通过网络管理系统或其他业务系统，调

用相关 API, 以实现灵活的备份/恢复功能。服务器 API 根据功能及操作对象, 可分为三类: 基础 API, 包括标准连接、存储设备识别和卷管理等; 存储管理 API, 包括存储设备加密、卷快照和复制、存储容量管理和负载均衡等; 介质数据管理 API, 包括数据缓冲优化、介质数据复制优化和 D2D2T 迁移等。

2) 虚拟介质池 (Virtual Media Pool) 技术

虚拟介质池, 将所有的介质服务器和所有介质服务器上的存储统一管理, 可以达到提高易用性、提高管理性的效果, 当介质空间不足时, 只要增加新的介质即可扩展总的存储空间, 虚拟介质池的性能会随着客户端数量的增加而不断提升。主要功能: 统一的路径访问、动态介质与自动漂移、全局空间配额与自动空间管理、介质无缝扩展。

虚拟介质池根据介质服务器的优先指数 (P_i) 选择介质, P_i 的计算方法如公式(1)所示:

$$P_i = \mu S_i / \sum S + (1 - \mu)(1 - T_i / \sum T) \quad (1)$$

其中, S_i 为第 i 个介质上的剩余空间, T_i 为第 i 个介质服务器上已绑定的任务数, $\mu \in [0, 1]$ 为调节系数, 若当前任务所需空间较大时, $\mu \rightarrow 1$ 。

3) 磁盘暂存技术与多层次的重复数据删除技术

磁盘暂存技术提供了一种快速备份/恢复方法, 首先在磁盘上创建备份镜像, 然后将镜像转移到另一种介质上。磁盘暂存以更快速的磁盘储存作为中间媒介, 大幅提升备份与复原程序的效能。因此, 更能缩短备份时间, 结合磁盘效能优点, 以及磁带长久保存可移植性优点。

重复数据删除技术主要包括前端的备份软件客户端的重复数据删除和后端基于硬件的虚拟磁带库重复数据删除技术。客户端重复数据删除, 在靠近数据源的位置进行, 采用了基于滑动分块 (Sliding Block) 方法的相同数据检测技术, 可以减少网络通信量和存储空间, 增加备份速度; 虚拟磁带库的重复数据删除采用了基于可变分块 (Content-defined Chunkin) 方法的检测技术, 以最小的存储开销识别和消除数据环境中的冗余数据, 减少存储的冗余数据, 最大化利用磁盘空间。

4) 系统日志分析与故障诊断报告

为便于管理员随时查看系统备份状态和结果, 将备份系统日志中的数据进行抽取, 并按管理员的需求进行多维分析, 将结果以图表或其它直观方式显示在

网络管理系统界面或 WEB 页面。

同时分析系统运行日志, 对故障点状态进行对比, 并进行聚类分析, 查找故障原因或故障出现时的环境特征, 以便实现系统的故障分析及预警, 帮助管理员更好的管理和维护系统。

2.5 系统备份策略

系统备份策略在主备份服务器上集中实施, 系统根据备份策略向逻辑数据云或物理数据云备份, 并且由虚拟介绍池自动选择最优介质进行备份, 系统管理员无须考虑存储介质的地理位置和状态。不同数据(文件)的备份策略如下所示:

1) 系统业务数据(文件)备份采用全备、差备与增量相结合的方式。备份策略如表 1 所示:

表 1 系统备份策略表

周日	周一	周二	周三	周四	周五	周六
F	I	I	C	I	C	I

说明: F 表示全备份; I 表示增量备份; C 表示差分备份

该策略充分运用了全备份、增量备份、差分备份的特点, 尽可能减少每次备份的数据量以提高备份速度。而且在任意时间点发生数据灾难后, 为恢复数据所需的备份 IMAGE 不超过三个, 保证了恢复的高效性。

2) 操作系统数据一经配置生成在一段时间内数据变化不会大, 因此其备份只有在操作系统有改动及系统被破坏时才会进行。对于操作系统及应用系统的备份采用每月分时段全备份, 在其余的时间我们采用增量和差分备份相结合的方式。

3) 数据库系统数据每周进行整个数据库全备份。这个备份是做数据库灾难恢复的基础。另外, 在每天做数据库的增量备份(即日志备份), 以实现数据库的时间点恢复。另外, 对于不同应用的表空间, 还可设定独立的备份策略以加强其安全与恢复的能力。

3 系统实施效果

通过该系统的实施, 浙江中烟实现了跨地区、大数据量的系统备份, 异步数据云和灵活的备份策略也使得备份效率得到很大的提高, 具体体现在:

1) 系统全备份(数据量约 22T) 时间由原先的 22 小时减少到 12.5 小时, 增量备份(数据量约 5T) 时间由 4 小时减少到约 2 小时, 恢复时间由原先的 2-3 小时减少到小于 30 分钟; 数据可用性从原先的 80% 提高到 99.99%; 而在运行成本上, 存储采购费用节约了

30%；IT 人员花在备份上的时间减少 50%。系统备份在不同数据量下的备份时间比较如图 6 所示。

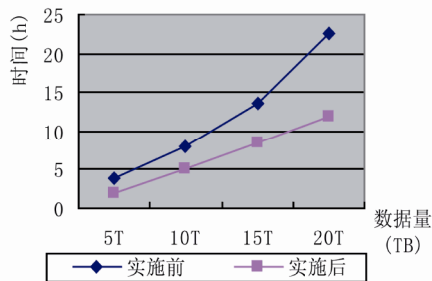


图 6 系统实施前后备份时间对比

2) 自动化、集中化的备份方式满足了系统管理员对系统文件、数据库数据的日常备份与恢复需求，且无须考虑备份介质的使用率和可用性问题，随时方便了解备份进度及状态，减少了系统管理员负担。

3) 数据备份和恢复时间大大降低，同时提高了备份的安全性和可靠性，从而大大增加了企业信息系统的安全性和抗灾害能力，也为后续的容灾系统提供技术基础。

4 结论

本文针对传统备份系统无法满足企业大规模、高效率 and 灵活性要求的现状，将云存储思想引入备份平台，提出了基于异步数据云的备份平台架构，以网络

交换层为核心，将逻辑数据云和物理数据云接入网络，实现稳定、高效、安全的海量数据备份，同时针对不同备份对象制定灵活的备份策略，提高了备份和恢复速度，通过在浙江中烟的实施，使其信息化管理工作效率显著提高，取得了明显效益，也为企业建设容灾、减灾系统提供了基础。

参考文献

- 1 任欣,李涛,胡晓勤.远程文件备份与恢复系统的设计与实现.计算机工程,2009,35(10):112-114.
- 2 韩塞北.网络存储主流技术比较及分析.长春师范学院学报(自然科学版),2010,29(6):38-41.
- 3 田萍芳,鲁宏伟,秦磊华.网络存储技术.计算机与数字工程,2004,32(2):38-41.
- 4 刘金芝,余丹,朱率率.一种新的云存储服务模型研究.计算机应用研究,2011,28(5):1869-1872.
- 5 高建秀,吴振新,孙硕.云存储在数字资源长期保存中的应用探讨.现代图书情报技术,2010,(6):1-6.
- 6 边根庆,高松,邵必林.面向分散式存储的云存储安全架构.西安交通大学学报,2011,45(4):41-45.
- 7 李翠侠.基于混合云的高校图书馆存储方案研究.图书馆学研究(理论版),2011(3):68-71.
- 8 刘国萍,谭国权,杨明川.基于云存储的在线备份安全技术研究.电信科学,2010(9):79-83.

(上接第 126 页)

MU 分类模型。

3 结语

本文提出 MU 分类模型，通过不断更新均值中心可以有效的对测试数据进行准确的分类，该算法不受样本分布概率模型限制，不用控制参数，使用简单，在此基础上又提出了 MU-MD 分类模型。MU 分类模型和 MU-MD 分类模型在训练样本较少或训练样本的均值离质心较远的情况，与 MD 和 NN 分类法相比较，其准确率较高。最后，通过实验比较，验证了本文提出的分类模型的有效性，以及在训练样本较少、训练样本的均值偏离质心较远及噪声情况下分类的优越性。

参考文献

- 1 Toth D, Aach T. Improved minimum distance classification

with Gaussian outlier detection for industrial inspection. Italy: Proc. of the 11th International Conference on Image Analysis and Processing. 2001.584-588.

- 2 Tan SB. An effective refinement strategy for KNN text classifier. Expert Systems with Applications, 2006,30(2): 290-298.
- 3 张孝飞,黄河燕.一种采用聚类技术改进的 KNN 文本分类方法.模式识别与人工智能,2009,(12).
- 4 任靖,李春平.最小距离分类器的改进算法——加权最小距离分类器.计算机应用,2005,25(5):992-994.
- 5 干正如,曾宪珪.基于自适应距离原理的自适应分类方法.江西理工大学学报,2007,(8).
- 6 Aeberhard S, Coomans D, Devel O. Comparative analysis of statistical pattern recognition methods in high dimensional setting. Pattern Recognition, 1994,27(8):1065-1077.