基于云存储技术的非结构化数据管理平台研究及实践

蓝玉蓉 张海静 肖伟 陈柏希
成都航空职业技术学院 四川 成都 610100

【摘 要】在高校数字化校园建设过程中，会产生大量数据信息，按照类型分为结构化数据和非结构化数据，非结构化数据包括存于 Word 文件中的电子教案、试题库，以 PDF 形式存在的教学课件、文献资料，以图片格式存在的示例图示，以及以视频格式存在的课堂实录、教学案例等。文章分析了高校非结构数据存储、管理的困境，提出由云服务提供商提供技术支持，搭建基于私有云存储技术的非结构化数据中心平台，从而实现学校非结构化数据集中存储、有效利用的目标。

【关键词】非结构化数据；云存储；私有云；资源平台

引言

随着现代计算机网络技术的高速发展和高校数字化校园建设，产生了大量的数据信息，数据作为一项资产，在提升教学信息化，管理信息化的能力和水平方面，起着重要作用。

数据按照类型分为结构化数据和非结构化数据，学生的学籍信息、教职工人事信息可以看作是结构化数据，可以用关系型数据库来表示，与结构化数据的规律性相比，非结构化数据不能用数字或统一的二维行列来表示，它们以文本、图片、音频、视频等形式呈现，无法用统一的、规范的概念来描述，根据互联网数据中心的一项调查显示，80%-90% 以上的数据属于非结构性数据，而且每年还以 63% 的速度增长，其中，视频、图片等多媒体格式的比重日益增加，又占到非结构性数据总量 70% 以上。在高等教育中，常见的非结构化数据包括存于 Word 文件中的电子教案、试题库，以 PDF 形式存在的教学课件、文献资料，以图片格式存在的示例图示、学校活动，以及以音频、视频格式存在的课堂实录、教学案例等。其中文本容量相对较小，对硬件没有特殊要求，但是图片、音频、尤其是视频占用存储容量较大，管理较为困难，因此在大数据时代，如何利用云存储技术保存好海量数据，提高平台的开放性，方便用户使用，是当前需要解决的问题。

一、高校非结构化数据现状

（一）非结构化数据增长速度快、数据量大

根据国际数据中心（IDC）预测，随着数据爆炸式增长，2018 年到 2025 年之间，全球产生的数据量将会从 33ZB 增长到 175ZB，而中国的数据圈正在迈向全球第一，随着联网人口持续上升，视频监控基础设施不断普及，中国数据圈增速最为迅速，2018 年，中国数据圈占全球数据圈的 23.4%，即 7.6ZB。预计到 2025 年将增至 48.6ZB，占全球数据圈的 27.8%，中国将成为全球最大的数据圈。就高校而言，随着最近几年计算机技术、网络技术的全面普及，数字化校园建设大力推广，学校的教学活动以及行政管理都开始通过计算机网络来实现，由此产生了大量的教学文件、课件资源等非结构化数据信息，同时早期以纸质或磁带为介质保存的教学资源和学校资料也被转化为数字资源加以保存，以成都航空职业技术学院电视台为例，早在多年前，就开始逐步将使用磁带保存的教学资源、历史视频转换为计算机数据格式存储，2012 年全面启用数字设备采集视频资源和图片以后，两者容量更是逐年递增（见图一），图片筛选一目了然，但大量视频以片段形式存在，长短不一，格式不同，按照素材采集要求，其中 80% 其他内容与有用的内容结合在一起，不能直接删除，依靠人工筛选保留有用素材，删除无用内容，费事耗力，工作量巨大。

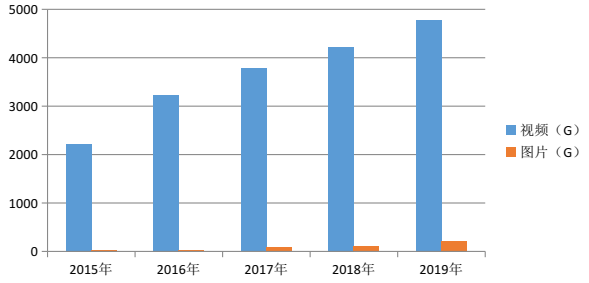

图 1 最近五年视频图片增量

（二）非结构化数据存储困难

面对快速增长的庞大的非结构化数据，储存和备份

是首先需要面临的问题。目前，高校中最常见的数据保存方法是各部门自行负责，教务处保存学生学籍信息、教师教学档案、学生工作部保存学生活动记录、保卫处保存监控视频、档案室统一管理学校文件、历史资料，大量非结构化数据孤立存在于不用部门，不同地点，由专职人员使用本部门业务系统进行存储与备份，基于不同应用构建的存储系统，最终导致数据重复、臃肿、分散，后续存储扩展和运维成本也比较高。

（三）非结构化数据流转困难，缺乏处理分析的技术手段

对于已经保存的非结构化数据，使用并处理它，其实是一项费力不讨好的工作，因为体量、网速、距离等因素的影响，非结构化数据的流转并不容易，更不用说被灵活的应用在数据分析和处理流程之中了，随着大数据时代的开启，各大科技公司已经意识到数据所带来的可观的经济价值，在商业方面，通过对用户行为数据属性的提取和分析，可以对目标客户进行广告精准投放，甚者改变用户的行为习惯，如果在教学中，也能对学生的学习内容、学习时间、学习习惯等行为数据进行提取和分析，就开展更有针对性地指导，从而提高教学效果。但是，目前大多数非结构化数据只能满足本部门查询，调用，其他业务部门如果想使用公开数据，必须联系专职人员，通过必要的业务流程，采用必须的硬件设备才能获取数据，耗时长，效率低，造成使用成本增加。

二、云存储技术概述

云存储技术属于云计算的一个分支，属于云计算的具体表现和应用，是指通过集群应用、网络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用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的一个系统（见图二）。

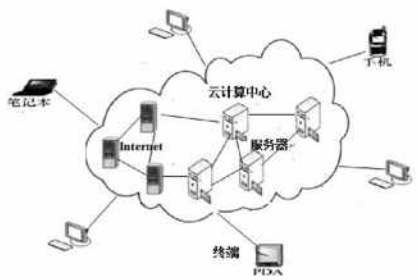

图 2 云存储管理架构图

云存储就是对数据进行存储、下载、共享处理，且不占用用户本身的计算机空间，也不必花钱去另外购买其他的存储设备，经济实惠，优势明显。二次访问操作也很方便，只要有网络，就能通过移动终端下载存储数据，比如进行课堂教学时，教师提前将课件、案例、资料上传到云空间，同学们可以提前下载预习，老师也可以在课堂上直接登录云空间访问资料，实现无 U 盘上课。这样既节省纸质资源，避免不必要的浪费，又可以做到长期保存，并随时更新。

（一）云存储技术类别

1. 公有云

公有云由互联网企业、电信营运商等第三方公司投资运营，面向所有互联网客户提供服务，每一个用户都可以自由选择，根据需要免费或付费租赁、购买云空间，独立使用和上传、下载云存储系统中的数据，而不需要承担基础设施建设成本和风险，极大节约了用户的资金投入。公有云通常在远离客户建筑物的地方托管，公众最为熟悉的就是百度云、阿里云、腾讯云、360 云盘、苹果 iCloud 等。

2. 私有云

私有云通常由第三方公司部署在企业或高校内部，单独为本单位提供服务，前期基础建设投入较大，由单位出资购买服务器及相关设备设施，本单位 IT 人员负责系统运行与维护，如果后期存储不够，需要扩容，也由单位支付扩容费用，但没有云服务费支出。私有云存储系统相对而言具有独立性、私密性、只限于内部使用，因此增加了数据安全性。

3. 混合云

混合云是指在搭建了私有云的基础上，又使用公有云服务，两者相结合，形成内部数据相互流动，混合型云存储具有上述两类云存储的特点，尤其当发生工作负荷快速波动时，混合云可以缓解负荷压力，用户可以获得最佳利益组合，具有弹性特质。

（二）私有云优势

对使用者来讲，云存储不是指具体一个或多个存储设备，而是用户数据访问服务，核心是应用软件与存储设备相结合，通过应用软件来实现存储设备向存储服务的转变。对企业或高校而言，在基础建设资金允许的情况下，部署私有云，对非结构化数据地存储与管理更为合理有利。

1. 确保数据安全性和持久性

目前公有云最大的问题是数据安全性无法得到绝对保证。近年来，互联网企业已多次出现由于运营维护困难，经营成本增高，导致直接关闭云服务，造成数据丢失；另一种情况是将增加的费用转嫁到客户头上，因为数据流转困难，迫使客户不得不增加自身成本继续接受服务，骑虎难下，给用户也造成极大的损失和麻烦；而且由于互联网开放的环境，给数据安全也造成极大的隐患，用户数据被分享，被泄露。私有云是为一个客户单独使用

而构建的，通常部署在高校防火墙内，使用环境相对封闭，使用客户相对单一，安全性可以做到自主控制。

2. 确保数据上传下载稳定性

使用公有云传输数据，既要取决于网络本身使用情况，还要取决于向第三方公司付费情况，在网络速度稳定的情况下，付费越多，下载速度越快，使用免费服务，则会出现上传容易下载难的现象，尤其是数据量大的情况下，数据传输缓慢，耗时费力；采用私有云技术，高校内网目前都是以千兆网速传输，通过分布式资源调度功能，实现本地存储资源的优化聚集，保障在各个时间节点内，都可以支持用户快速上传下载数据，满足了数据传输过程的稳定性。

3. 保证硬件设施有效整合

采用云存储技术统一存储数据，统一管理，减少了高校内部多部门各自为政，自主采购维护计算机硬件设施的情况，云平台在搭建初期和开始提供服务后，硬件设备必然由专门部门和专业人员负责维护和管理，后期其他部门使用，只是用到云服务，而不需要对硬件设备负责，当存储容量不足时，也只需要专业人员考虑扩展容量，非专业人员无需关注云存储后台运行与维护情况，确保设备利用率提高，得到有效整合。

三、私有云存储技术在非结构化数据存储与管理中实践

私有云存储技术在实施过程中有两种途径，一种是自主设计，自主开发管理平台，自行搭建系统，另一种则是与第三方公司合作，由第三方公司提供成熟的管理平台，自购服务器集群搭建系统。根据学校的实际情况和经费预算，我们采用了第二种模式，搭建了以视频、图片储存、在线浏览、下载为主的私有云平台。

（一）平台设计思路

非结构化数据中有 70% 为视频、音频、图片等多媒体信息，学校在长期办学中也积累了大量的视频和图片，搭建校内云存储平台，可以方便使用者自行查阅和下载，减少工作人员重复劳动，提高工作效率，因此平台需要提供无感化（无需输入账号密码）、可视化操作，满足视频在线流畅播放，考虑到教学需求，平台还需要与 BB 教学平台整合对接，在不改变师生使用习惯的前提下，为 BB 平台提供底层强大的流媒体支撑，解决 BB 平台视频播放卡顿的问题。

平台包含三层架构：物理存储层、基础管理层、应用接口层（见图三）。

物理存储层是平台最基础的部分，主要解决异构存储设备的整合，构建统一的存储中心，实现非结构化数据集中、高效、安全存储和统一管理。

基础管理层是平台最核心的部分，主要负责处理应用接入层的请求和物理存储层的非结构数据实体调度，该层基于分布式计算框架搭建，根据自身需求部署多个计算节点，满足用户对可靠性和效率的需求，包含上传、转码、检索、备份等服务。

应用接口层是最灵活多变的部分，提供标准的 REST OPEN APT 接口，供其他系统和服务接入数据中心，对非结构化数据进行相应操作，例如文件检索、视频转码、图片下载、多终端播放、统计分析等。

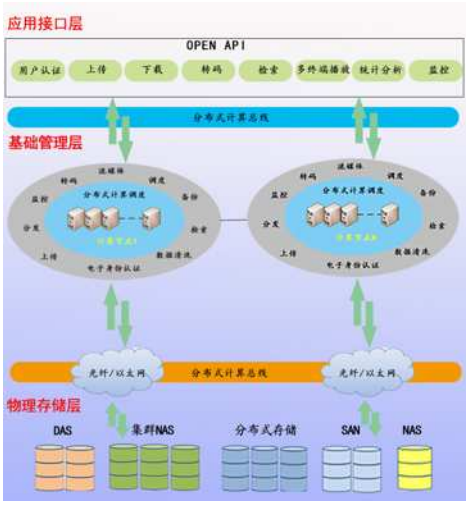

图 3 非结构化数据中心平台架构图

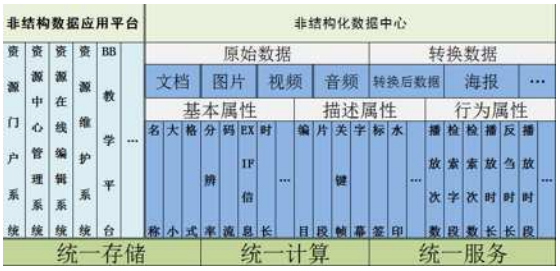

图 4 平台结构

平台包含了非结构化数据应用平台和非结构化数据中心两大模块，数据应用平台包括资源门户系统、资源中心管理系统、资源维护系统等基础系统应用，也可以按需求接入其它应用，数据中心主要用于对视频、图片、文档等非结构数据进行统一存储、统一管理，主要包括原始数据、转换数据（转码后数据、海报数据等）、基本属性（名称、大小、格式、分辨率、码流、EXIF 信息、时长等）、描述属性（编目、片都、关键帧、附件、字幕、标签、水印等）、行为属性（播放次数、检索字段、检索次数、播放时长、播放时段、反刍时长等）。

（二）平台应用

在资源门户系统主页上，既可以按照图片、视频、课件、教学文件等格式设置若干入口，也可以按照部门

名称设置站点入口，频道数量可以根据需求创建，以成都航空职业技术学院校园电视台为例，根据开设栏目创建了《成航新闻》《成航大舞台》《校园拍客》等 9 个频道，管理员根据视频内容分门别类，把不同的视频上传到不同频道中，任意用户可以通过校园内网进入资源平台，无需注册登录，直接访问，即可浏览或下载公开视频或图片，给用户带来极大便利。另外主界面还包括公告、推荐、排行榜等多个模块，信息分类明确，针对点击较多的资源，系统采用热点推荐的方式进行数据推广共享，数据在上传过程中，自动截取海报，以可视化的方式呈现在网页中，使用户可以轻松实现对非结构化数据的实时浏览和下载。

图 5 成都航院校园电视台平台

在线看视频，对服务器系统是一大挑战，针对大流量视频浏览，非结构化数据中心在保留源文件的情况下，专门提供了转码服务，包含 CPU 转码、GPU 转码、单机多进程转码、集群转码、切片转码、定时转码、多码流转码，无论采用哪种转码方式，系统均在后台自动完成，通过分发服务可以将部分或全部数据分发到一个或多个异构存储上；热点数据通过分发服务按照策略分发到高 I/O 的服务器 SSD 硬盘上，结合负责均匀服务，流媒体服务，通过交换机网口绑定或者平台多网卡调度服务，共同提供低成本大开发的应用支撑。

（三）平台管理

非结构化数据平台实施两级管理模式，提供了系统管理员和站点管理员两种角色，系统管理员处于上一级，负责整个整个平台硬件与软件维护，包括站点创办，用户管理等等，保障平台平稳运行，站点管理员处于下一级，负责对所属站点数据进行上传、管理与维护。系统通过后台设置，提供不用类别用户权限设置功能，可实现不同栏目、不用数据的指定使用。

四、结语

基于私有云存储技术的非结构化数据管理平台的研究，为非结构化数据管理平台的建立提供了依据，通过非结构化数据中心对非结构化数据进行统一存储、统一计算、统一服务，把非结构化数据中心的数据同结构化数据相结合，提供大数据应用，为管理型的数字校园向服务型的智慧校园转型提供了有力支撑。

【参考文献】

[1] 陶皖 , 李钧 , 李丞龙等 . 云计算与大数据 [M]. 西安 : 西安电子科技大学出版社 ,2017,12.

[2] IDC: 年均增速 30%,2025 年中国将以 48.6ZB 领跑全球数据圈 https://blog.csdn.net/weixin_33794672/article/details/89572032.

[3] 陶皖 , 李钧 , 李丞龙等 . 云计算与大数据 [M]. 西安 : 西安电子科技大学出版社 ,2017,54–60.

[4] 许豪 , 邱雅 , 曹蕾等 . 云计算导论 [M]. 西安 : 西安电子科技大学出版社 ,2015,47–52.

[5] 钟庆 . 非结构化数据平台在教学实践中应用——以传奇数字资源云服务平台为例 [J]. 现代教育技术 ,2018(7):77.

【基金来源】

本文系 2017 年度成都航空职业技术学院自然科学基金研究课题“基于云存储技术的非结构化数据管理与应用研究”（061771）成果之一。