

高校数据资源池建设的模型设计与运行机制研究

黄传峰, 高渡新, 陈铭轩

(南京工程学院 经济管理学院, 南京 211167)

摘要: 数据时代的来临, 高校内学生的多维异构数据被大量采集, 致使众多高校面临着数据来源分散、异构严重的问题, 进而在校园信息管理中引发数据冗余量大、利用率低等现象, 导致无法进行有效的数据挖掘和数据的关联利用, 造成“信息孤岛”现象与“烟囱式”管理问题; 该文章设计了高校数据资源池的统一管理模型, 通过数据采集、存储管理与应用访问的闭环整合, 利用数据虚拟化等相关理念和技术, 通过非关系型数据库的特性融合异构数据, 实现资源池内数据的过度融合, 以选课系统为例的实验结果表明: 该模型有效解决了各种异构数据的融合管理与集成应用问题, 实现了高校统一的数据资源管理与综合应用的平台化运行, 能够更好地支持数据时代下的高校学生服务、教学科研管理与后勤保障工作。

关键词: 数据资源池; 高校; 模型设计; 运行机制

Model Design and Operation Mechanism of Data Resource Pool Construction in Universities

Huang Chuanfeng, Shang Duxin, Chen Mingxuan

(School of Economics and Management, Nanjing Institute of Technology, Nanjing 210093, China)

Abstract: With the advent of the data age, the multi-dimensional heterogeneous data of students in colleges and universities has been collected in large quantities, which has caused many colleges and universities to face the problem of scattered data sources and serious heterogeneity, which leads to large data redundancy and low utilization rate in campus information management. As a result, effective data mining and data utilization cannot be performed, resulting in “information island” phenomenon and “chimney type” management problem. This paper designs a unified management model for university data resource pools. Through data collection, storage management and closed-loop integration of application access, using data virtualization and other related concepts and technologies, the heterogeneous data is integrated through the characteristics of non-relational databases to realize resources. The over-integration of data in the pool, the experimental results of the course selection system show that the model effectively solves the problem of integration management and integration of heterogeneous data, and realizes the unified platform operation of data resource management and comprehensive application in colleges and universities. It can better support college student service, teaching and research management and logistics support in the data age.

Keywords: data resource pool; universities; model design; operating mechanism

0 引言

大数据时代下, 信息与数据的利用服务于各行各业, 深刻影响着人们的工作、生活, 并推动了高校信息化建设 2.0 时代的变革与发展。与此同时, 非一线行业如高校也暴露出了脱产的弊端, 即高校所使用到的信息数据资源价值远远小于其实际蕴含的价值, 因此我们不仅可以将硬件资源信息化, 同时考虑将高校内的数据也整合为虚拟化资源

池供数据仓库进行过滤提炼, 并用于数据挖掘和决策分析。而数据资源池技术作为一种将数据、服务器, 存储、网络、都整合成一个虚拟的资源池 (如图 1), 这样应用软件需要的资源可以在资源池里抓取, 不仅能够提升高校的服务器资源的利用率、数据可用性, 同时也可以指导高校管理做出科学决策。



图 1 高校内虚拟资源池整合

收稿日期: 2019-04-29; 修回日期: 2019-06-07。

基金项目: 南京工程学院科研创新基金重大项目 (CKJA201706); 江苏省大学生实践创新训练计划重点项目 (201811276004Z); 南京工程学院大学生科技创新基金项目 (TP20180005)。

作者简介: 黄传峰 (1969-), 男, 江苏沛县人, 博士后, 教授, 主要从事信息系统和大数据管理方向的研究。

虚拟资源池作为实现融合基础设施结构与数据的关键要素,是共享服务器、存储和网络的集合,能够根据应用程序的要求更快地进行重新配置,从而使管理员能够比以往更容易、更快捷地支持业务需求的变化。在云计算环境下,资源不再是分散的硬件,而是将物理服务器经过整合之后,形成一个或多个逻辑上的虚拟资源池,共享包括计算、存储、网络资源。资源池可以委派对集群资源的控制权,在使用资源池划分集群内的所有资源时,其优势非常明显。首先可以创建多个资源池作为主机或集群的直接子级,并对它们进行配置。然后可向其他个人或组织委派对资源池的控制权^[1]。在高校信息化环境下的大数据统一(管理)资源池平台是基于特定用户大量的行为数据在统一管理资源池平台中进行数据分析、利用及可视化处理用于为学校的管理带来智能化的改变和效率的提升。

1 相关研究

目前对资源池的应用,众多学者和业内人士已经在多方面展开了研究。葛中泽^[2]等提出对资源池数据访问模式的探讨,通过对传统模式和资源池模式的测试对比,针对提高服务器端的资源利用率、提高系统的性能为目的,提出了基于资源池模式原件设计的一般方法,并作为引例详细阐述了图书管理信息系统中的资源和资源池的设计,以及资源池的管理策略和管理线程。该研究在资源池的运行模式建设中有较高的利用价值,在物理层方面服务器的资源利用率确实是一个瓶颈,对于高校来讲,服务器的高利用率意味着可以减少硬件的堆叠,降低资金投入,减少运维成本同时带来更好的用户体验。在资源池的建设方面,刘波等^[3]分析了基于私有云的资源池建设要点。针对数据流量的爆发式增长,为提升用户体验、业务效率引入 IT 支撑系统资源池部署以及云计算和大数据技术的应用,提出需架构需集中化、平台化、安全又灵活的 IT 支撑私有云架构。其中针对 IT 支撑系统分散,组网架构技术标准不统一等现状采用明确资源池对上层应用提供通用接口的要求并不断进行优化。在高校的资源池建设中也可采用同类方法处理不同业务系统提供的信息数据从而加以整合形成可利用的信息资源池,用以打破高校目前面临的信息“孤岛”问题。沈锐^[4]等提出基于资源池的资源聚合机制,由于系统的动态性和自治性特点非常突出,系统服务质量很难得到保证,而该应用机制将所需资源合成资源池,并在运行过程中即使部分节点由于系统故障、超载等原因中途退出任务处理,用户的服务请求仍能得到处理和呼应。该机制的应用极大缓解了机器故障带来的不确定性和紧张感,其将需求虚拟化和本文提出的数据虚拟化相类似,有效地提高了系统的服务质量。刘振恒等^[5]设计了基于开源 KVM 虚拟胡资源池以及动态的 JBoss 负载均衡技术的框架结构,实现了一个能够按需分配的资源池,提高了高校内数字化资源动态。崔嘉^[6]探讨了云计算的资源池管理,较详细地针

对数据资源、计算资源等进行了分析,对私有云资源池管理具有一定借鉴作用。沈志豪、梅峰等^[7]针对设备资源利用率不高,同时设备维护管理成本持续增加研究了基于虚拟化的资源池技术。

高校内部的信息系统建设本身对用户体验、储存能力、智能化有着较高要求。通过引入资源池建设与“数据闭环”的模型设计与运行机制,以解决高校信息资源利用率低、用户体验差、业务系统分散、信息孤岛等问题。

2 校园数据资源与虚拟化

数据资源虚拟化是资源池运行的前提,也是智能化校园建设的主题和目标,当今大数据(智能)的浪潮已经进入各类行业,在大数据时代和高校信息化建设 2.0 时代的双重发展下,资源池建设作为提升校园管理智能化、个性化的新途径,也对高校的教育管理产生了深远的影响。高校作为高等教育的人才培养中心,在许多特定时期及日常工作中都会产生大量与学生相关的信息数据,这些数据也为高校带来了重大机遇和艰巨的挑战。高校数据呈现多样化、异构化、多源且跨行业整合的大规模数据,特别是半结构化和非结构化的数据,必须解决数据收集、数据存储、数据分析、数据管理、数据应用等多方面问题^[8-9]。

资源池的模型对处理大数据方面具有提高应用性能等先天优势,带来的高效率并行处理(massively parallel processing, MPP)逐渐成为高校治理体系和管理能力现代化水平的关键因素。因此将资源池技术积极纳入校园管理中,使高校在校内进行各种活动时所产生的各种数据成为高校制定管理方案的“决策源”,从而进行更为精确的信息决策有着更加深远的意义。然而许多高校的信息中心通常将重点放在已经可视化的结构化大数据中,对学生日常活动及特殊教学工作所产生的非结构化数据没有引起足够的重视。对管理者来说全面了解学生在高校中的真实需求和学生的基本生活状态,提升高校的教学质量,全面素质教育教育和推动高校信息化建设 2.0 时代的发展,将数据整合在资源池中加以应用,在校园管理中就显得尤为重要。

在数据虚拟化环节中,由于高校内获得教学信息的途径多样,且数据比较分散,存在数据重复收集,汇聚精确信息难,二级部门收集后保存时间短,同时由于信息基础架构模式没有很好的规划,很多系统数据仅仅为了解决某些特定的业务而存在,没有针对性的做集中化及数据聚合,本应有紧密联系的相关管理系统,实际上却没有相应的关联,很大程度上影响了共享资源池的数据利用,造成难以对数据综合利用。与此同时大多数在高校数据储存部分未对数据进行分类和虚拟化,而是直接在物理设备上存储,也就必然存在成本高,应用效率低,运维难度大等问题。

对师生而言,资源池的建设也可以为用户提供更加个性化和智能化的服务,也为改善传统教育管理模式的种种

弊端提出了新的解决方法。由于高校大多数信息业务紧紧围绕学生的学业活动, 具有较强的峰谷负载特征。但是目前多数高校的信息平台通常是按照当时峰值负载来设计的, 这导致业务空闲期大量资源的浪费, 同时在业务高峰时无法满足用户需求, 也给信息中心带来不便。

在高校数据资源池模式建设与运行机制方面, 由于大部分高校对数据价值的认识还停留在表面, 重硬件堆叠, 轻数据深度利用, 从而缺乏对数据资源的统一规划和深入分析, 从而导致各部门在数据资源管理方面缺乏有效的管理机制, 导致数据资源效能不高, 因此在数据资源池建设方面需要考虑高校内复杂的业务逻辑及流程、处理大量学生数据、相关子系统之间的业务系统融合等。

3 高校大数据资源池模型应用

该模型基于高校的业务现状, 利用数据资源虚拟化, 结合资源池技术和理念, 提出了智能化的闭环数据资源池信息框架, 如图 2 所示。

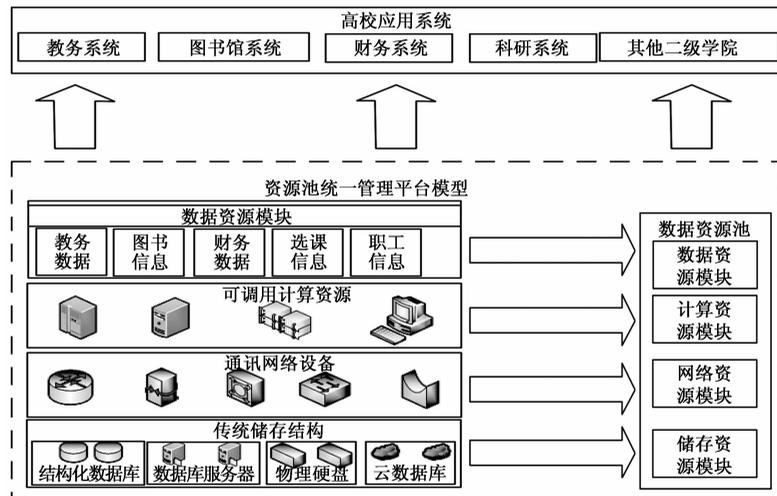


图 2 高校数据资源池模型

高校资源池应用模型由资源池和管理运营平台组成, 资源池主要由计算资源模块、网络资源模块和存储资源模块, 存储资源模块又分为云资源模块和本地资源模块。在该平台模型中并不直接存储数据, 也不进行数据挖掘和分析, 正如物理硬件层面的虚拟化一样, 而是对数据资源进行重新整理和抽象虚拟化, 再结合外部提供的资源进行业务需求处理。

3.1 计算资源模块

计算资源 (computing resource) 模块对资源池的支撑关键在于底层物理元件的异构融合, 将物理逻辑结构上相对独立的硬件实现虚拟逻辑上的整合, 然后重新定义, 不再受限于单台物理服务器的计算界限, 使如中央处理器 (CPU) 等硬件设备成为动态管理的对象, 实现 $1+1>2$ 的效果, 从而为资源池的运行提供充沛的计算能力。计算资

源模块由高校现有的服务器资源以及云服务器计算资源融合得到, 通过多台真是物理服务器和虚拟服务器组成服务器群, 服务器虚拟技术对物理服务器进行重新整合, 实现了已有资源的高效利用, 同时提升了整体计算性能和系统管理能力^[10]。随着信息化的不断发展, 高校内的办公教学系统不断增多, 功能需求不断提高, 各教学管理中心和二级学院之间的数据中心造成极大的计算资源浪费。与此同时, 当高校有新功能和业务并发增加时, 物理资源的增加能力和成本额度也十分受限, 但虚拟化的计算资源模块具有动态自增长和收缩的特点, 很容易满足新业务和高并发的计算资源需求。

3.2 网络资源模块

网络资源 (network resource) 模块是数据资源池抽象后完成数据处理和资源虚拟的匹配, 网络虚拟资源模块包括了网络节点、链路、接口等基本要素, 校园网内包含两套网络系统, 包括局域网和广域网, 通过 NAT 接入外网。校园局域网并存着教务系统、科研服务、财务系统, 在这之间又包含对学生和教师开放的与你专用网络 (VPN) 服务。高校内多种网络应用服务并存于相同的网络环境里, 因此网络的高可用性、扩展性和高速传输就显得尤为重要。对应资源池的网络结构采用网络控制和应用控制, 通过优先级策略分配贷款, 结合身份验证和安全隔离技术实现高效的网络服务资源分配。

3.3 数据资源模块

数据资源 (data resources) 模块是资源池虚拟化的核心模块, 在该模块中实现虚拟数据对象的控制, 进行对数据重新整合、确立权威数据、进行抽象和提取等一系列操作, 并为上层提供统一数据接口。数据虚拟对象是储存资源模块的数据进行融合, 抽象出新的数据对象^[11], 提供给资源池平台统一使用。通过储存资源控制器达到屏蔽底层硬件的物理异构, 对资源池而言储存资源模块的数据是一片“水”, 具有透明性、一致性强且集中, 对计算机系统而言, “资源池化”后的数据相当于一个高可塑性的“电资源”, 相对的同构化且具有高度的可用性, 通过资源池模型的处理直接产出“数据产品”。对用户而言, 可以只关心处理后的数据结果, 在终端完成不同院系间、不同系统间的高效产出工作。实现了数据便捷、一致、集中的结果, 进而可以完全不用理会数据异构型带来的不兼容问题, 不同院系间的信息孤岛、各个系统间数据壁垒高引发数据不通用的问题。

同时该层不会开放数据库的存储和提取权限, 也无法通过如结构化查询语言 (SQL) 等进行系统操作或事先格式化的操作, 而是通过模型化的算法、数据接口函数、认为定义等方式进行汇聚处理和虚拟化。该模块通过在软件

维度上的数据虚拟化也就意味着完全摆脱了传统意义上的数据分析、数据管理和数据应用,不再受限于人工+管理系统的形式,而是引入人工智能模型,通过算法对数据进行加工和处理。在数据分析方面由于全部数据基于全部统一标准,因此可以为数据仓库提供更为快速和准确的数据,而不用再考虑数据源的底层异构和融合问题。数据虚拟化是对数据表象的一种升华,其背后的真实数据控制器决定了融合后数据的可用性和便利性。每一种虚拟数据对象都配备有相应的虚拟数据对象控制器,负责提取数据汇聚层下方的真实数据格式,然后封装为对象的虚拟数据对象进入数据汇聚层,提供给上层统一规范后的数据进行使用,与此同时上层传递过来的数据再通过数据控制器解封传给数据汇聚层进行真实的数据存储和传输。例如学生对象(student)包含基本信息(student_ID、student_num)等多个主键和外键,而通过数据控制器会提供具体方法产生一个查询多个数据库后集成的表单,这样就避免了因高校内多个业务系统存在交集的数据而缺乏权威标准,同时又弥补了单一数据库信息不全的缺点。如遇到数据产生交集且不一致时,权威数据可以通过分析业务流的数据流来确定。如学生基本信息来源于学生入学时在教务系统录入产生,而后分发给其他系统,因此从业务流上分析学生的基本信息如遇到自其他系统中或者产生变动应以教务系统的数据为权威。同样的交集数据情况也可以通过这种方法来确定权威数据。

3.4 储存资源模块

储存资源(storage resources)模块是虚拟资源池模型真实数据存储的核心,该模块围绕真实数据信息提供存储系统,对数据进行备份和容灾,包含本地物理存储和云服务器存储,该模块为资源池提供原始数据,这些数据可以是结构化、半结构化或者非结构化的,可以来自关系数据库也可以来自非关系数据库。通过储存资源控制器达到屏蔽底层硬件的物理异构,对服务器而言处理的是同构的统一数据,会大大增加储存器的储存能力,提高利用效率。

4 基于高校数据资源池的运行机制

高效资源池的应用改变了传统IT系统的资源配置模式,使得IT系统的资源配置更灵活、高效。资源池在高校多子系统的环境下实现资源、管理、服务方面的协同演进构建智能化、高效化、动态化的资源服务平台。

资源池管理运营体系与实际生产环境的结合应用机制受到更为广泛的关注和重视。通常高校内系统建设方式导致系统资源无法共享、响应业务需求慢的现状,资源池通过主机、存储等虚拟化技术建立跨系统的共享资源池,提供资源的动态分配,具有高可用性的处理能力。上述文章针对资源池的建设和运营体系进行了深入研究,其中设计研究方法为高校资源池的建设提供了较好的借鉴。但高校

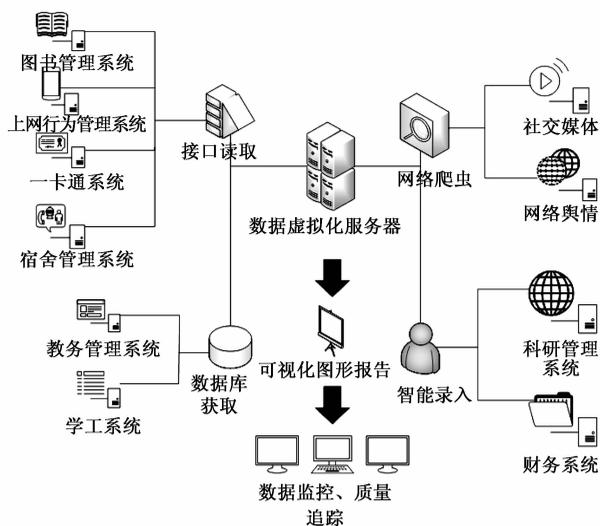


图3 高校数据资源池平台的运行架构

内数据往往错综复杂,各业务部门系统之间的数据冗余度高、无统一规范,需要更具针对性的数据资源池应用。

在该资源池模型应用机制下,把高校内所有子系统和业务系统和学生端软件(如选课系统、教务系统、图书馆系统、信息办业务软件)部署在该资源池上,所有的应用均在该资源池客户端上运行,实现C端的融合,后端的运行区分。高校的学生数据资源由“分散”到“集中”,有效地避免了信息的孤岛化,模块化,管理费用和建设费用也势必减少,高校管理的过程中为了更加智能化人性化的管理学生,以学生的需求为出发点,大数据时代下了解学生需求的最有效的途径就是通过学校的管理系统手机与学生相关的数据,提取和分析其中的潜在价值,进而分析学生的真实诉求。基于虚拟化软件简历资源池统一管理平台,实现资源的智能调配、资源池数据安全、子系统接入等应用。该应用机制实现对服务器集群、计算资源模块,储存模块、数据模块等基础资源环境的统一管理、实现基于用户优先级差别化性能管理和性能分配,提升对用户行为控制的效率,通过智能化的流程提供服务。

5 业务运行结果分析

通过分析高校内的业务现状,以高校选课系统为试点,结合业务环境遇到的问题利用资源池技术加持对比得出一些实验结论。引入资源池模型后,使选课系统的后端结构由后端程序+冷数据库转变为后端程序+Redis+虚拟资源池,由于资源池是使用虚拟同构数据,数据存储在虚拟资源池中,所以比一般存储在硬盘中的数据库有IO的先天优势,同时由于Redis是使用Hash表的内存数据库,所以查询效率在最优情况下为 $O(1)$,极少情况到达最坏情况 $O(n)$,可以在高并发量下显示出优势。

表 1 Jemter 测试数据

Label	# samples	Average	Min	Max	Std. Dev	Error/%	Received / (KB/s)	Sent/ (KB/s)	Throughput/s	Avg. Bytes
Login	200	1 338	81	10 722	1 790.97	0.00	2.94	3.59	17.1	176.0
Getc	200	8 893	77	20 851	5 304.11	7.50	13.21	2.44	9.4	1 432.1
Select	200	365	62	3 415	671.98	0.00	1.61	2.55	9.4	175.0
Total	600	3 532	62	20 851	5 012.46	2.5	16.29	6.91	28.1	594.4

后端程序在选课前使教务系统中与学工系统冷数据库中的异构数据载入资源池, 通过资源池的数据虚拟化模块整合选课所需关键数据, 并且通过读取前端返回数据进行选课 (如图 4), 在访问量减少时和结束时, 将 Redis 数据迁移到资源池中, 从而完成选课数据的虚拟同构化。



图 4 选课系统前端界面

在传统的服务器上扩展, 将服务器分为后端程序服务器与资源池服务器, 并且对这两块服务器进行集群处理, 用以提升查询效率。对于资源池应用, 设置协调服务器, 动态监测服务器的状态, 对 CPU 设置阈值, 将数据处理尽可能平均地分配给多个资源池服务器, 并且在资源不足时申请新的云服务器, 将新申请的服务器使用虚拟资源池服务器的镜像, 同时加入集群。对于虚拟资源池, 将需要同步的数据 (比如选课的课程余量) 设置在单一服务器, 将不相干数据设置在多个服务器 (比如学生学号与密码, 学生所对应的空闲时间段, 课程), 将部分可分割数据设置在多个服务器 (学生学号对应的空闲时间段, 课程, 将各个不同的学院布置在多个服务器上)。数据资源池对于后端程序和数据库, 相当于私有云数据库, 由于有专门的数据和程序服务器, 使得查询和传输后端数据的资源分摊在多个设备上, 同时得益于资源池数据的统一性和随时动态扩展的优势, 可以在数据量再次扩大时对资源池进行升级或添加新的数据服务器, 以适应更大的并发。本项目使用 vmware server 进行服务器的虚拟化, 便捷地将服务器虚拟出宿主操作系统, 以最大程度凸显资源池的虚拟环境。对于

后端程序服务器与数据库服务器进行单独测试, 使用 C# 或其他本地化的语言写测试程序, 利用 jemter 测试 Web API 的承载并发能力 (如表 1), 可以减少浏览器对于网页渲染对测试后端程序服务器与数据资源池的服务器的影响, 以更真实地请求完成资源池效能测试。

6 结语

高校数据资源池的建设利用数据资源池技术、虚拟化技术, 在原有硬件虚拟化的基础之上实现了高校内异构数据的高度集成, 为信息中心对大数据的应用提供了统一的源数据环境。资源池平台的搭建为高校提供了一个信息统一管理平台, 实现了对数据动态、高效地灵活调度, 同时和物力资源进行智能弹性组合, 很好地解决了数据差异大带来的数据资源利用率低的问题, 必将为学生、老师带来更大的便利, 也为高校的信息化发展提供了良好的数据发展空间。

参考文献:

- [1] 张涛. 资源池模式及其扩展策略应用研究 [J]. 电脑知识与技术, 2011, 7 (7): 1578-1580.
- [2] 葛中泽. 基于资源池的数据访问模式的探讨 [J]. 科学技术与工程, 2012, 12 (33): 9066-9070.
- [3] 刘波, 黎炜, 陈保福. 私有云资源池的建设要点分析 [J]. 电信科学, 2016 (S1): 209-214.
- [4] 樊华, 沈锐, 王戟. 虚拟计算环境中基于资源池的资源聚合机制 [J]. 计算机工程与科学, 2009, 31 (3): 122-127.
- [5] 刘振恒, 李天工. 云环境下高校数字化资源自适应调度框架研究 [J]. 计算机科学, 2012, 39 (s2): 123-125.
- [6] 崔嘉. 基于云计算的资源池管理系统的研究 [J]. 自动化技术与应用, 2017, 36 (5): 19-21.
- [7] 沈志豪, 梅峰, 蒋鸿城, 等. 资源池虚拟化技术的应用研究 [J]. 电力信息与通信技术, 2015, 13 (3): 39-44.
- [8] 赵国锋, 葛丹凤. 数据虚拟化研究综述 [J]. 重庆邮电大学学报 (自然科学版), 2016, 28 (4): 494-502.
- [9] 罗伟雄, 时东晓, 刘岚, 等. 数据虚拟化平台的设计与实现 [J]. 计算机应用, 2017 (a02): 225-228.
- [10] 罗伟雄, 曾纪霞, 时东晓, 等. 基于数据虚拟化技术的高校云数据中心构建 [J]. 软件导刊, 2018, 17 (3): 194-197.
- [11] 葛苏慧, 梁宏涛, 房正华. 高校共享数据中心虚拟化技术的架构 [J]. 计算机技术与发展, 2014 (4): 174-177.