高校数据中心数据资产归集管理研究

史进[1,2]　李剑云[3]　江晓峰[4]

1.普元信息技术股份有限公司；2.上海交通大学电子信息与电气工程学院；

3.中通服软件科技有限公司；4.上海仪电鑫森科技发展有限公司

[摘　要]数据作为一种新型生产要素，其重要性正在日益凸显。随着高校信息化建设和智慧校园的飞速发展，产生的数据呈爆发式增长，高校大多选择建立数据中心解决数据的归集和使用问题，本文阐述高校数据中心建设和数据资产管理的研究现状以及存在问题，并详细阐述了数据资产归集管理应包含的内容，包括数据模板管理、数据集成管理、数据质量管理以及数据分类管理，最后对高校的数据资产归集管理进行了总结，为高校数据中心数据资产管理体系的构建提供了依据。

[关键词]高校数据中心；数据资产；归集管理；数据集成管理

【DOI】10.12252/j.issn.2096-627X.2020.02.926

一、引言

学校是培养人才的摇篮，高校教育在其中扮演非常重要的角色，承担着为国家输送高素质创新型人才、构建创新型人才培养体系、引领国家级研究课题的重任，为了适应国家数字经济战略的发展，高校的信息化建设一直处于高速发展的阶段，高校持续投入资金进行网络环境、基础硬件设施和软件设施，全国各高校的信息化系统也日趋完善，而高校由于存在大量的教学、科研类或者行为特征等实时数据，越来越多的高校选择建设集约化的数据中心，用来归集、处理日常产生的海量数据。虽然有了数据中心，但由于数据的归集和处理需要跨部门协同工作，且数据治理和数据资产管理体系的理论起步较晚，目前缺乏针对高校数据中心的数据资产管理的体系研究。

二、研究现状

（一）智慧校园及教育信息化

在国家层面，相继发布了针对智慧校园，教育信息化建设的相关标准以及指导性文件。国家标准化委员会在2018年发布了《智慧校园总体框架》国家标准，教育部2018年印发了《教育信息化2.0行动计划》[1]，描绘了数字时代教育信息化的阶段特征，提出2022年基本实现“两全、两高、一大”的发展目标，2019年，国务院出台《中国教育现代化2035》[2]，其中第八项“加快信息化时代教育变革”明确提出建设智能化校园，统筹建设一体化、智能化教学、管理和服务平台。各省市也颁布高校智慧校园方面的管理办法，但这些标准及管理办法更多聚焦于信息化、数字化建设，对于高校的数据资产如何管理以及使用缺乏相关标准和法规依据。

（二）高校数据中心建设研究现状

在高校数据中心建设的研究方面，学者多聚焦于如何建设高校的数据中心平台，如刘勇在《基于云计算的高校数据中心设计和实现》[3]中提出基于云计算的高校数据中心设计，从IT资源、云平台、高可用等维度，对计算资源、存储资源、网络资源及其虚拟化进行了设计。刘嫚在《高校数据中心建设及数据治理探究与时间——以中国地质大学（北京）为例》[4]中提出数据中心的功能架构以及数据治理体系，包括规章制度建设、一数一源，一源多用、定期数据质量检查、数据可视化等方面。倪宇斌在《高校数据中心平台原型设计研究》[5]中提出，数据中心的平台应包含管理主题分析、教学主题分析、学生主题分析、决策分析以及数据上报的功能。

（三）高校数据管理研究现状

在高校数据管理研究方面，学者多聚焦于高校数据治理，少部分学者聚焦于校园的数据资产管理的研究。其中，高嵩在《高校教育数据里的应用研究》[6]中提出教学数据应该从管理转变为治理，异构数据的治理、教学数据治理系统的构建等方面进行了探讨。陈方方在《高校数据治理机制探索及成效》中提出运用信息技术工具按照一定标准对大数据进行采集、清洗、存储和删除等处理过程。罗军锋在《基于高校的数据资产管理模型》中提出数据治理主要包括数据的标准化、数据的清洗、数据的交换以及数据集成技术。蒲天银在《高校教育数据治理问题及对策》中提出需要对高校数据进行分类，并提出数据治理管理机制模型和教育数据治理举措。郑苑在《教育信息化背景下高校数据治理研究》中提出高校数据治理的框架设计以及数据治理的管理体系，包括数据标准管理，数据流程管理和数据资产管理，最后提出了数据管理的实施过程和效果。

吴学刚在《校园数据资产的沉淀》中提出利用大数据手段关注学生的微观表现，利用大数据手段整合优质教育资源。邹蕾在《高校数据资产管理现状和对策研究》中提出树立“数据即资产”的理念，做好顶层设计，构建学校的全局数据中心，设计标准数据接口，建立基于全局数据库的数据监测分析平台。

（四）存在的问题

目前校园数据资产管理处于起步阶段，还存在诸多问题，比如数据孤岛的问题、数据质量不高、数据资产价值有待挖掘等问题。

1.数据孤岛现象普遍存在。由于顶层设计的缺失，高校在不同时期建立了很多业务系统，各个业务系统间并未打通，存在数据孤岛的现象，为后续的数据分析及使用带来很大的困难。

2.数据质量整体不高。高校业务系统由不同部门牵头建设，数据源不同，并未形成统一的数据标准，也没有规范化的数据归集、处置流程，数据质量较低，存在大量数据缺失、数据不规范的现象，为后期的数据分析应用带来很大的局限性。而且高校存在大量非结构化数据，该类数据在收集时没有统一标准，只能依靠邮件或者即时通信工具流转，数据散乱，还增加数据风险。

3.数据资产价值有待挖掘。虽然高校拥有海量数据，包括实时数据和非实时数据，但目前高校数据管理的重点在于管理各业务系统，鲜有针对数据的分析和利用，缺乏数据可以为科研、为领导决策提供依据的功能，也缺乏整体的数据协同服务，数据资产没有充分发挥价值，缺乏对数据资产价值的挖掘，从而也无法利用数据分析降低学校运营成本，提高运营效率。

三、数据资产的归集管理

为了解决高校数据资产管理存在问题，首先需要解决数据的归集问题，通过构建高校数据的归集管理体系解决数据资产的归集问题。高校数据资产的归集管理一般可包括数据模板管理、数据集成管理、数据质量管理、数据分类管理四大部分。其中数据的模板管理是前提，数据的集成管理是归集的主要手

段，数据的质量管理是数据可以有效使用的检核措施，数据的分类管理可以为数据的标签化使用提供有力支撑。

（一）数据模板管理

高校的数据资产一般包含结构化数据和非结构化数据，结构化数据一般存在各个业务系统的数据库中，如学生信息、教务信息、教师信息、设备信息等。非结构化数据指文档类、图片类等以非结构化形式存在于系统中的数据，如学生照片、获奖证书电子版、教学课件等。

针对结构化数据，需要建立命名规则，梳理业务系统表，绘制关联ER图，确定字段含义等。针对非结构化数据，需要梳理非结构化数据的相关关键信息，制定对应规范，如文件命名规则，存放路径，文档编写人等。

高校数据中心针对不同业务系统的归集，需要建立归集的规则，如归集的主键和时间戳管理、归集频率管理、数据接口管理等。归集的时间戳一般包括技术时间戳，业务时间戳。

（二）数据集成管理

高校数据存在大量的结构化以及非机构化数据，不同类型数据集成方式各不相同。针对非结构化数据，一般采用逻辑入湖的方式。针对结构化数据，一般采用物理入湖或入库的方式，即通常所说的ETL过程，主要包括数据抽取、数据清洗、数据转换和数据加载4个步骤。

1.数据抽取。数据抽取的作用是从高校不同时期建设的各业务系统中抽取数据到应用层数据库的过程，数据的抽取需要在调研阶段做大量工作，并建设面向后期应用的各主题库、专题库如科研主题库、教学主题库、学生主题库等。常用的数据抽取方式包括全量抽取和增量抽取。

2.数据清洗。数据清洗的作用是处理原始数据中的脏数据。由于原始数据中有可能存在着大量的脏数据，脏数据产生原因各异，如采集缺失产生的脏数据、数据保存不当产生的脏数据等原因，可以利用有关技术如数理统计、数据挖掘或预定义的数据清洗规则将脏数据转化成满足数据质量要求的数据。脏数据主要类型包括不完整数据、错误数据和重复数据三大类。常用检测方式是通过SQL设定清洗规则，批量查找出脏数据，然后反馈至学校各个业务部门，补齐以及修改相关数据。

3.数据转换。数据转换的作用是将不同数据源的数据按照一定的标准或者常用的业务计算规则整合成统一的数据形式并进行存储的过程。如每个学生在学校都应该拥有唯一的ID号，但实际操作中则不是如此，学生在录取管理系统会分配唯一ID号，学籍管理系统分配唯一ID号，就业指导管理系统分配唯一ID号，但同一学生在不同系统中的号均不相同。又如学校收集的学生填写个人信息中，民族的填写可能有“满族”“满”“manzu”“man”等多种填写方式，这类问题需要通过标准化代码映射的方式进行统一的标准化处理，通过数据转换功能将数据统一转换成唯一编码，如使用身份证号作为学生的唯一标识符，将填写的民族转换成统一的表示方式等。

转换过程可以采用数据库存储过程转换或者高级语言转换。数据库存储过程转换即使用SQL开发存储过程完成转换作业，高级语言转换包含了常用的开发C/C++/JAVA等程序对抽取的数据进行预处理。

4.数据加载。数据加载是将转换作业生成的数据插入目标数据库，一般加载作业只需要使用INSERT或者LOAD的方式导入目标表即可。根据抽取作业的数据抽取方式的不同（全量、增量），加载的方式也会有所不同。一般包括文件加载、落地加载及不落地加载等方式。

（三）数据质量管理

数据质量管理是保障数据资产质量的重要环节，质量管理会设定质量检核规则，如值域规则、业务规则或非空质量规则等。通过质量校验规则，对数据进行初步的校验，筛选出异常数据，常规的校验规则，包括完整性校验、重复性校验、有效性校验等。完整性校验主要校验集成数据是否完整，是否包含空数据、空记录数等；重复性校验主要校验数据在单表中是不是存在重复记录数；有效性校验主要校验数据是否在业务指定的值域范围之内。

（四）数据分类管理

高校的数据分类规则主要是对归集的数据做初步梳理，从满足数据梳理要求的角度出发。核心是从主体、业务活动等维度建立一整套适用高校的分类体系模型，对各个数据打上对应的标签，以满足数据使用的需求。分类的主要目的是能够从业务角度上较宽泛的划分主题业务域，便于下一步数据模型设计。

关于根据教育数据的分类，结合高校数据的实际情况，将高校的数据按照数据主体、业务活动、采集技术三个维度进行分类，其中，数据主体可以分为学生、教职工、家长和社会三个类别，业务活动包含科研类数据、教学类数据、社会服务类数据、行政管理类数据、资源类数据，采集技术包含物联网感知类数据、应用系统采集数据，见表1。通过以上分类基本可以将所有的校园数据进行归类。有了对应的分类依据，数据中心需要对各类数据打上对应的标签，以便于数据的统一管控以及场景应用。

每个分类都必须有对应的编码规则，定义统一的编码和名称，建立系统之间编码与统一编码的对应关系，一般是以表的形式进行存放，也可以在资产管理平台中统一维护。

四、结论

通过解决高校数据中心的数据资产归集问题，可以为高校数据分析提供高质量的数据，最大化发挥数据作为一种要素的价值，为智慧校园和教育现代化建设提供决策支持，尽管目前数据资产的价值没有得到充分利用，但未来通过构建完善的高校数据资产管理体系，驱动管理模式的创新，可以切实提高数据资产价值，为高校教育改革带来质的飞跃。

参考文献：

[1]教育部关于印发《教育信息化2.0行动计划》的通知.教技〔2018〕6号.

[2]中共中央、国务院印发《中国教育现代化2035》.

[3]刘勇，陈云峰，郝璐瑶.基于云计算的高校数中心设计与实现[J].微型电脑应用.2019,37(10):13-17.

[4]刘嫚，于磊.高校数据中心建设及数据治理探究与时间——以中国地质大学（北京）为例[J].中国教育信息化.2019,(19):79-82.

[5]倪宇斌.高校数据中心平台原型设计研究[J].软件.2019,43(2):70-72.

[6]高嵩，赵卓，袁艳朝.高校教育数据治理的应用研究[J].科技创新导报.2019, (11):200-201.

作者简介：史进（1983-），男，博士，职称：高级工程师。上海交通大学电子信息与电气工程学院和普元信息技术股份有限公司联合培养博士后，主要研究方向为政府及能源行业信息化项目管理、数据资产管理、运营、评估与交易。