文章类型：论文|刊号（ISSN）：2972-4236(P) / 2972-4244(O)

基于多源异构数据环境下的高校数据中台研究

胡博
西安培华学院 信息中心
DOI:10.12238/acair.v2i2.7388

[摘　要] 随着信息技术的快速发展,高校在经历了教育信息化“1.0时代”的建设和发展,即将转向以流程服务、数据服务为核心的教育信息化“2.0时代”,高校数据中台作为数据汇聚、处理、分析和服务的核心平台,其研究具有重要的现实意义。本文针对基于多源异构数据环境下高校数据中台的建设展开研究,分析了多源异构数据环境下高校数据中台的内涵与特点,探讨了高校数据中台的关键技术,构建了高校数据中台架构,并通过具体案例验证了高校数据中台在提升数据质量和优化决策支持方面的显著效果。
[关键词] 数据中台；多源异构数据；全数据链体系；数据治理；数据开放
中图分类号：N37　文献标识码：A

Research on university data center based on multi-source heterogeneous data environment

Bo Hu
Xi 'an Peihua University Information Technology Center

[Abstract] With the rapid development of information technology, colleges and universities have experienced the construction and development of education informatization "1.0 era", and are about to turn to the education informatization "2.0 era" with process service and data service as the core. As the core platform of data aggregation, processing, analysis and service, the research of college data center has important practical significance. This paper conducts research on the construction of university data middle offices in a multi-source and heterogeneous data environment, analyzes the connotation and characteristics of university data middle offices in such an environment, discusses the key technologies of university data middle offices, constructs the architecture of university data middle offices, and validates the significant effects of university data middle offices in improving data quality and optimizing decision support through specific cases.
[Key words] data center; Multi-source heterogeneous data; Full data link system; Data governance; Data opening

引言

2018年4月,中华人民共和国教育部印发了《教育信息化2.0行动计划》教技〔2018〕6号文件,意味着进入2.0阶段后,高校的管理改革步伐必然大大加快,数据的综合利用将成为高校改革的主要动力。首先,高校将推动数字化转型,深度融入人才培养过程,并利用人工智能、虚拟现实等技术进行教学环境、过程和方法的革新。其次,将加强对数据、流程、服务、应用等方面的标准化治理,建立全面而精细的标准体系,使数字化转型更加有序和高效。同时,高校将越来越多地采用云架构,提升数字化基础能力、运维能力和服务能力。此外,教育部深入推进国家教育数字化战略行动,通过建设国家智慧教育平台、慕课建设和应用等方式,促进教育质量的提升和教育公平。数字化变革正推动着教育资源的共享,无论是基础教育、高等教育还是职业教育,都在通过数字化平台提供更优质的教育资源和服务。教师数字素养和胜任力的提升被视为教育数字化的重要组成部分。政府支持与政策导向也将大力发展数字教育,并将其作为推进教育现代化和强化高质量发展基础支撑的重要策略。这些趋势共同指向一个更加数字化、智能化和个性化的教育未来。

1 亟待解决的问题

数据资产是高校教育信息化建设的核心价值之一,可以通过完成海量数据的采集、整合、梳理及存储等工作,从中挖掘有价值的数据,但是在真正的数据分析开展的时候,由于数据存在多源异构的环境影响,各单位往往会存在着担忧。即数据的可靠性不强、数据质量问题严重,很大程度上影响了预期的效果,基础性数据出现的问题,严重限制了信息化水平的提升。目前还存在以下亟待解决的问题：

1.1数据不可知

作为终端的使用者无法获知系统平台中有哪些数据,更无法获知这些数据和业务之间的关系,如何有效的获取这些数据。

1.2数据不可控

由于没有统一的数据标准,导致数据难以集成和调用;由于没有数据质量控制,导致海量的数据因其质量过低而无法被应用；由于没有能统一管理的平台,导致数据管理流程不可控。

1.3数据不可取

作为终端的使用者不能便捷自助地拿到数据,导致业务分析的需求难以被快速满足。

1.4数据不可联

实际工作中产生的大量结构化和非结构化数据,受制于其知识体系的差异,在数据和知识之间无法建立有效的关联以及转换,导致无法对数据开展探索和挖掘。

1.5数据不可用

无法基于数据说话,数据的应用和服务价值难以体现。

2 主要研究内容

在多源异构数据中,“多源”通常是指在获取数据时,原始数据的多样化的来源。例如在高校中除了各类业务系统产生的数据,还包括各类设备日志、保存在个人PC上的文档、表格、图片和音视频等文件,都属于数据的来源。“异构”则通常是指获取到的数据,在数据的内容或者模型结构上存在的差异性。常见的异构数据可以分为两大类,即以数据库为代表的结构化数据,和以日志文件、图片或音视频为代表的非结构化数据。

国内最早关于“中台”概念,是由阿里巴巴集团首创,同时形成了独特的“中台战略”。“数据中台”则是中台战略的核心技术,一种战略选择和组织形式,是依据终端用户特有的业务模式和组织架构,通过有形的产品和实施方法论支撑,构建一套持续不断把数据变成资产并服务于业务的机制。

基于互联网行业先进的数据中台理念,结合高校教育信息化建设实际的现状,从多源异构数据环境中,将数据集中、数据梳理,数据共享、数据资产管理、数据应用五个层面以数据中台的形式进行整体规划构架,就是本文所探讨的基于多源异构数据环境下的高校数据中台。

多源异构数据环境下的高校数据中台的建设不是一蹴而就的事情,在建设路线上应打破原有“点状应用孤立建设”的状况,优先治理底层基础数据、夯实底层基础数据,再辅以标准化的管理方法,才能确保数据质量的提升,从而转化成为高质量的高校数据资产,形成可迭代更新的全数据链生态体系。因此建设高校数据中台需要从以下几个方面进行：

2.1完成数据中台基础数据的治理

制定统一信息标准,规范数据来源,采用统一数据录入(产生)标准,保障数据的规范性。提供统一的、规范的数据计算、数据共享接口,杜绝数据冗余,实现数据的“全”、“统”、“通”。高校各业务系统产生的多源异构数据,在经过数据治理过程后,可以实现统一标准存储以及和数据质量的提升,既为数据平台的数据分析模型提供了高质量的数据输入和多维的分析角度,同时又为后期数据分析打好了基础。

教育部2012版信息标准(JY/T1006—2012)规定了高校管理信息系统管理的基本标准,包括基本体系结构、库元数据结构、高校管理基础数据元素等内容。该标准适用于高校在建设业务管理系统时,对数据规范和数据结构进行定义。数据标准设计涵盖数据标准制定、审核与公布。数据标准通常是在数据治理的开始阶段,由负责数据标准管理的部门进行设计制定的,并以学校公文的形式正式公布并严格执行。随着信息化的不断发展,当数据标准发生更新迭代时,还需要管理部门及时将最新版本的数据标准发布更新,以便各业务管理平台和数据基础平台,可以按照最权威的、最新的标准对系统平台中的数据进行定义、对数据表进行设计。数据标准设计作为成果数据存储的格式规范,需要由数据中台系统统一管理。

2.2形成数据中台全数据链体系

全数据链体系是指对数据的生命周期和使用痕迹形成全面的管理和自动化追溯跟踪,即数据从产生到消亡的全过程可记录,数据从源头到终端所有转换过程可追踪,实现链条式、链路式数据管理和应用体系。并且能够形成全员参与的数据资产管理体系,让所有使用数据(包括接口数据、辅助决策图表数据、个人数据服务)的程序或人员,都能够了解数据来源,随时可反馈数据质量问题,让数据治理过程全员可参与,所有问题可追溯,数据可以持久化、常态化治理,数据质量及数据能力可螺旋式不断提升。满足数据管理在“便捷、高效、可控”这几个方面的需求。

常见的数据仓库通常基于hadoop分布式架构结合oracle数据库的形式。采用企业级大数据产品Cloudera的各种高性能组件,如Spark、Sqoop、Flume、Kafka等作为处理引擎,实现对全量数据的海量存储、高效计算、挖掘分析。

整个数据仓库的建设都依照数据标准进行建模,使用建模工具结合数据标准规定的分类和格式规范,生成相应的数据仓库结构,再采集高校的各种有价值数据,按照质量要求进行清洗治理,按照数据标准的格式进行建模,利用大数据基础技术架构进行存储,形成全量数据仓库。同时,对重要状态数据进行历史数据积累,形成全生命周期数据资源体系。

全量数据仓库将用来支持流程服务、数据调用、交换共享、大数据分析、精准管理、科学决策等事务。对采集的结构化数据提供标准化存储服务。要求数据的组织方式和存储结构符合高校的校级数据标准相关要求。

2.3开展基于数据中台的数据应用建设

经过数据中台基础数据治理和数据中台全数据链体系的建设,底层技术架构、数据标准和数据质量提升工作完成后,数据分析所需的准备工作即告完成,接下来就是针对现有的数据来源、数据分析、数据梳理的成果,创建能够开展数据挖掘和分析的应用场景,设计与其相关的业务指标或算法模型,进行数据分析和数据挖掘。为使用者和开发者提供符合标准的数据支持,

降低开发成本，提高开发效能。常见的数据应用包括数据开放平台、智能数据门户、数据资产可视化系统。

(1)数据开放平台。数据开放平台作为数据应用的核心系统之一，需要基于前期完成的数据中台建设的基础上，为高校用户提供数据开放服务，引入第三方应用开发，构建高校大数据应用生态，提供灵活的管理模块，方便数据的存储、计算、标准的定义、发布、审核及使用审计。具有高负载和海量数据处理能力，可以基于多种类型数据源进行异构数据计算。同时提供完善的数据安全机制，对数据源进行隔离保护，保障数据的读取安全和存储安全。

(2)数据资产门户。对于数据中台应用，数据资产门户也属于核心系统之一，通常情况下在数据开放平台中会包含简单的数据门户，但这种数据门户一般只具备数据或接口的查询，以及简单的维护功能。因此数据资产门户作为展示数据资产目录的重要服务，可以通过其监控数据中台的数据状态，查询或调用数据中台的数据，成为面向数据管理者、业务部门、高校师生、软件开发单位的数据资产的供给平台。

(3)数据资产可视化系统。数据资产可视化系统作为业务数据、流程，工作状态呈现，以及交互的应用端，负责为决策部门提供数据支撑。通过可视化的交互界面，实现数据展示和多维度分析，助力决策部门进行正确决策。数据中台的建立使得数据成为决策层的重要依据，数据中台在前期完成了数据的采集、统计和分析，而这些数据成果则可以通过数据中台应用进行呈现，数据资产可视化系统作为数据中台成果的呈现平台，也被称为数据中台的最后一公里。

3 结束语

数据中台为高校的信息化建设提供的较为完善的解决方案，可以有效的解决在多源异构数据环境下高校数据的采集、存储、治理、分析，以及最为核心的数据开发和数据服务等任务，为高校提供数据集中、数据治理、数据共享、数据资产管理和数据应用开发的服务。详见如下：

(1)数据集中。实现对高校主数据共享的扩充，实现对业务数据的集成、线下数据的集成、同时实现对当前数据分析应用涉及到的日志数据、物联网数据进行集中，形成全量的数据资产体系。

(2)数据治理。深度调研高校各业务系统数据使用、共享情况，结合教育部教育行业标准，制定符合高校实际情况的数据标准，根据标准对已经完成集中的数据进行质量清洗，形成数据资产，结合统一数据仓库管理平台形成知识库，可以对数据标准、元数据、主数据以及数据质量进行管理维护，让数据治理工作可持续化进行。

(3)数据共享。对标准数据仓库中的数据，通过统一开放平台进行统一的管理和对外开放，规范数据审批、应用管理、接口管理等工作流程，确保数据使用规范，数据使用安全。

(4)数据资产管理。通过数据门户对数据资产进行全生命周期的管理，同时了解高校数据资产、了解数据共享交互情况、了解数据质量情况，结合质量监督、纠错进度监督等功能，督促各业务系统对数据质量进行提升，加强各业务系统数据管理意识。

(5)数据应用开发。利用各类数据资产挖掘数据价值，建设数据分析应用，通过数据辅助领导决策，给终端用户提供更好的信息化服务。

[基金资助]

西安培华学院2021年校级科研项目(PHKT2135)。

[参考文献]

[1]杨宗凯.教育信息化2.0:颠覆与创新[J].中国教育网络,2018,(01):18-19.

[2]翟雪松,楚肖燕,张紫徽,等.基于中台架构的教育信息化数字治理研究[J].电化教育研究,2021,42(06):40-46.

[3]白雪伟,杨疆,王涛.关于企业级智慧中台的构建方法研究[J].长江信息通信,2021,34(09):207-209.

[4]马晓玲,朱丽娟,吴永和,等.教育数据中台系统模型及其应用研究[J].现代教育技术,2021,31(11):63-71.

[5]张洁,许建宏,肖伟.关于数据中台建设思路的探讨[J].邮电设计技术,2021,(08):74-79.

[6]李金旭,吕书林.高校大数据平台建设研究[J].电脑知识与技术,2017,13(16):13-14.

作者简介：

胡博(1982--),男,汉族,青海西宁人,硕士,西安培华学院信息中心副主任,研究方向：数据与大数据的分析挖掘,云计算及云平台应用,数据管理,智慧化教育。