基于高校数据中台环境下数据治理的研究

胡博

西安培华学院 信息中心

DOI:10.12238/acair.v2i2.7387

[摘要] 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”,通过以数据资源,更是以数据资产为中心,构建各类数据分析应用场景。本文围绕着高校1.0时代积累的海量数据,探讨基于数据中台来创建统一的和标准化的数据管理平台,将高校海量的数据归入其中进行管理,打破数据的壁垒,使用数据标准对数据进行管控、明确数据权责、提升数据质量,提升教育信息化系统平台的管理能力。

[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: C37 **文献标识码:** A

Research on data governance based on university data center environment

Bo Hu

Xi'an Peihua College Information Center

[Abstract] After experiencing the construction and development of the "1.0 era" of education informatization, colleges and universities are about to turn to the "2.0 era" of education informatization with process service and data service as the core. By taking data resources, especially data assets as the center, various data analysis application scenarios are built. Focusing on the massive multi-source heterogeneous data accumulated in universities in the 1.0 era, this paper discusses the creation of a unified and standardized data management platform based on the data center, which will incorporate the massive multi-source heterogeneous data in universities for management, break the barriers of multi-source heterogeneous data, use data standards to control data, clarify data rights and responsibilities, and improve data quality. Improve the management ability of education information system platform.

[Key words] data center; Multi-source heterogeneous data; Full data link system; Data governance; Data opening

引言

高校信息化建设历经数字化时代、信息化时代积累,已经基本完成了校园主干网络设备硬件的建设,并且根据高校业务体系,建设了一批业务管理平台和应用系统,完成了高校校园网基础业务的信息化建设。随着这些业务系统和设备的启用,产生了海量的各类数据,这些数据中蕴含了巨大的价值,记录了校内业务群体的行为过程及结果。但是高校现有的数据基础架构很难对这些跨平台、跨系统、跨技术、跨架构的数据,进行采集、存储、处理和分析,更不要说对高校的改革和决策提供支撑。因此急需创建统一的、标准化的数据管理平台,将高校海量的数据归入其中进行管理,打破数据的壁垒,以统一标准对这些数据进行管控,实现统一数据标准、消除数据孤岛、完成数据集中、创建全域数据仓库、解决数据协调问题、统一数据发布、明确数据权责、提升数据质量,提升教育信息化系统平台的管理能力、服

务能力和智能能力,同时从多个维度为高校提供数据分析和挖掘的支持。

1 高校数据建设发展现状

目前大多数高校已经基本完成了各业务部门的业务系统的建设,同时也建设了统一身份认证、统一门户、数据共享库等公共系统的建设,高校的教育信息化发展已向2.0时代进发。高校教育信息化的改革将以数据和大数据为驱动,构建未来高等教育的新形态,而作为建设基石的数据,目前还存在以下亟待解决的问题:

1.1 实现了对主数据的管理,但是无法形成完整的数据资产体系

经过前期信息化的建设,高校已经完成传统的“线下办公”到“线上办公”的转变,以解决各业务部门工作效率低下问题为目标,随着这些业务系统以及硬件设备的运行和使用,系统产生

和积累越来越多的数据,这些数据目前很难进行统一管理,导致大多数高校一直无法形成完整的数据资产体系,数据的使用存在很大的问题,也无法为大数据分析应用等提供有效的支持。

1.2数据标准未能进行统一,各系统数据孤岛现象严重

相当多的高校当前已经建设了相应的数据标准,但是由于在前期建设时,受到技术差异的限制,数据很难做到统一规划,数据标准无法有效落地,数据的存储是否按照数据标准进行存储,都没有有效验证为以后数据的互联互通埋下隐患。

1.3数据权责划分不明确,数据唯一性无法保证

在高校已形成的数据环境下,很多同样的数据是由不同的业务系统产生,如果没有明确数据权责,导致数据来源不唯一,面对不同的数据无法保证哪个业务系统产生的数据是权威而准确的,这就对后期使用数据造成很大的困难。

2 亟待解决的问题

数据资产是高校教育信息化建设的核心,但由于数据存在多源异构的环境影响,数据可靠性不强、数据质量低下、基础性数据出现的问题,严重制约信息化水平的提升。本文将主要探讨解决以下几个问题:

2.1统一标准,消除数据孤岛

结合部标、国标以及学校数据的实际使用情况,在学校原有的标准的基础上,将其他部门的数据进行整合,通过建立数据中心库,将数据整合到统一的中心库下,使数据具备关联分析的条件,消除数据孤岛。

2.2完成数据集中,建设全量数仓

通过对学校业务系统产生的业务数据、线下数据、网络设备安全设备产生的网络数据,以及物联网产生的设备数据、互联网产生的舆情数据进行统一的集中治理,生成我校全量数据资产,通过多维度多类型的数据实现更加精准的数据分析和应用。

2.3解决数据协调问题

数据整合完成之后,由原先“点对点”的数据共享模式变成“集中式共享”模式。所有需要数据的应用不再需要找业务部门协调数据,只需要由信息中心进行统一安排协调即可。业务系统厂商也不再重复收费的机会。将原先复杂混乱的接口全部集中统一管理之后,管理方面将更加规范和条理化,不需要进行复杂的关系协调和费用协商。

2.4统一数据发布

通过建设统一的数据交换平台,将所有数据共享转变为线上完成,实现自动化、高频化的数据统一发布,规范了各个应用的调用行为,并能对每个应用关联的接口、数据项、字段、调用频度进行管理和实时监控。

2.5明确数据权责

通过对业务流程的整体梳理,明确各部门数据供需关系,并对标准中涉及使每一项数据进行部门权责划分与确认,确保每个部门清楚自己需要维护和负责的数据,确保每一条数据都有了明确的来源。

2.6提升数据质量

检查学校各类存在的数据问题,通过技术手段实现数据各类质量问题的处理,提升学校数据质量,并形成良好的数据治理生态,持续化改善学校数据质量,逐步实现高质量的数据资产。

3 基于数据中台的数据治理

“数据中台”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。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高校数据中台下的数据治理。

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。因此需要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。在数据治理的开始阶段由负责数据标准管理的部门设计制定学校的数据标准,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,以学校公文形式正式公布,当数据标准发生更新迭代时,要及时将发布更新最新版本的数据标准,以便各业务管理平台和数据基础平台,能够进行相应的更新。数据治理工作的主要内容如下:

3.1数据调研

对学校的部门组织架构、管理机制、数据管理现状进行调研了解。对各管理部门的业务内容、组织机构、管理信息系统、后台数据库信息、数据内容和形态、供需关系、交换共享要求、数据质量问题、可引入的标准资源等进行调研了解并输出数据现状报告。

3.2校级标准制定

遵守国家强制标准、参考教育部推荐标准和其他学校数据标准、引入学校已有部分合格标准(含元数据模型和代码集),兼顾各个标准之间的兼容性、一致性以及标准的可扩展性,并匹配XX大学的管理业务特性,确定数据子集的分类方式,建设和完善XX大学校级数据标准。制定学校自定义的编码规范并给出数据分类编码规格说明书,输出一套符合学校实际的《XX大学数据标准》(数据子集、标准代码集、编码规范)。

3.3数据识别

根据数据调研的结果,罗列所有已知的数据来源和分布,记录访问账号和入口,查看实际的数据内容。根据上述数据治理范围目标,从数据来源中识别出目标数据所在的库、表、字段、格式等。

3.4数据采集

利用部署好的各种软件进行数据采集。其中:

(a)数据库:利用ETL软件采集(b)表格数据:利用线下表格数据处理工具软件采集和处理。如果项目中还需要采集互联网

数据,则可部署爬虫类软件进行采集。其中数据库和表格数据存储在临时库共享/集中库中,等待进行清洗治理。日志数据的采集后存储在Hadoop集群的HDFS分布式文件系统中。

3.5 质量检查

制定数据质量规则(如非空、唯一性、长度、取值范围、枚举范围、关联一致性等)。根据每个字段应有的质量属性将正确的质量规则绑定到字段上。执行质量检查,将不符合质量规则的记录识别出来,形成数据质量报告,以便评估数据质量状况、定位有问题的数据,便于下一步修正处理。

3.6 清洗转换和质量提升

根据质量报告的内容,确定质量问题的内容、程度,并定位到具体记录。这部分工作是数据治理的核心工作,是整个过程中最复杂、最耗时的阶段。大致上,质量问题分为2大类:

(a)结构性质量问题:如代码集定义、表达格式、数据单位不一致等问题。这类问题可以通过批量的、规则化的转换处理即可转换成符合标准的数据。这类问题,将通过ETL的转换规则实现质量提升。(b)内容性质量问题:如数据缺失、内容错误等问题。一般情况下,这类问题无法通过简单的批量转换修复,因此需要将问题数据和质量报告提交到数据对应的负责部门,由部门进行核实、采集、填报后形成有效数据。数据修复后,需要记录在业务管理系统中,通过再次采集入库,直至质量检查合格。

3.7 成果数据生成

将各种清洗转换完成、符合质量标准的数据导入到之前已经建模完成的表格中,即形成了符合校标的、质量达标的成果数据,即“标准化数据仓库”。这是数据治理项目最重要的成果之一。

3.8 数据封装发布

合格的成果数据由统一数据开放平台进行存储,并通过API接口、数据库访问、表格文件这3种方式向各个应用系统、数据

使用单位提供数据调用服务和数据共享交换服务。

4 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。实现对高校主数据共享的扩充,对业务数据的集成、线下数据的集成、制定符合高校实际情况的数据标准,让数据治理工作可持续化进行,确保数据使用规范,数据使用安全,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,给终端用户提供更好的信息化服务。

[参考文献]

- [1]杨宗凯,吴砥,郑旭东.教育信息化2.0:新时代信息技术变革教育的关键历史跃迁[J].教育研究,2018,39(4):16-22.
- [2]王祎琚,陈晓军.一种基于数据管理服务的一体化平台设计[J].中国科技信息,2023,(20):79-81.
- [3]李岳梦,熊文剑.区块链技术提高中台能力的思考[J].信息通信技术与政策,2021,47(9):91-96.
- [4]徐望,董剑峰.基于数据中台和服务中台的智慧校园框架设计与应用[J].中国教育信息化(高教职教),2021,(8):76-78.
- [5]虞振峰.高校教育数据中台建设的探索与实践[J].科学咨询(科技·管理),2021,(07):103-104.
- [6]Qiang Chou,Wei Fan,Jianming Zhang.A Reinforcement Learning Model for Virtual Machines Consolidation in Cloud Data Center[J].Proceedings of 2021 6th International Conference on Automation, Control and Robo,2021.07.

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。