

基于数据共享的高校数据融合策略研究^①

张文德¹ 杨雪² 谢建琳³

¹ (福州大学信息管理研究所, 福建福州, 350108) ² (福州大学经济与管理学院, 福建福州, 350108)

³ (福州大学图书馆, 福建福州, 350108)

摘 要 [目的] 构建基于数据共享的高校数据融合整体框架, 提出促进高校数据融合的策略。[方法] 通过分析目前数据融合过程中数据标准、数据冲突、数据同步方面的现状, 发现其存在的问题, 构建基于数据共享的高校数据融合整体框架。[结论] 高校数据融合过程中存在数据标准不规范、数据同步技术运用不科学、数据冲突与数据工作人员不专业的问题。制定数据标准、采用科学的数据同步技术、制定适当的数据融合规则及培养专业数据工作人员可以促进高校数据融合, 实现高校数据共享, 推动智慧校园的建设。

关键词 数据融合; 数据同步技术; 数据标准; 数据共享

Research on University Data Fusion Strategy Based on Data Sharing

Zhang Wende¹ Yang Xue² Xie Jianlin³

¹ (Institute of Information Management of Fuzhou University, Fuzhou, Fujian, 350108, China) ² (School of Economics and Management of Fuzhou University, Fuzhou, Fujian, 350108, China) ³ (Library of Fuzhou University, Fuzhou, Fujian, 350108, China)

Abstract [Objective] To construct the overall framework of university data fusion based on data sharing, this paper puts forward the strategies to promote university data fusion. [Methods] By analyzing the current situation of data standards, data conflicts and data synchronization technology in the process of data fusion, the existing problems were found, and the overall framework of university data fusion based on data sharing was constructed. [Conclusions] There are some problems in the process of data fusion in colleges

①本文系 2021 年度福建省中青年教師科研項目 (信息化专项) 基金項目“面向服務的数据融合研究”(項目編號: JAT211008) 的研究成果之一。

and universities, such as nonstandard data standards, unscientific application of data synchronization technology, data conflicts and unprofessional data workers. Formulating data standards, adopting scientific data synchronization technology, formulating appropriate data fusion rules and training professional data staff can promote university data fusion, realize university data sharing and promote the construction of smart campus.

Keywords Data Fusion; Data Synchronization Technology; Data Standards; Data Sharing

1 引言

随着信息技术的高速发展,智慧校园建设如火如荼,这为教育信息化建设奠定了良好的基础。智慧校园建设的规模不断扩张,大部分学校的各个部门都建立了信息服务平台来收集师生的个人信息,为师生提供便捷的信息服务,大幅度提高了高校的管理水平,基本实现了学校管理与教育业务流程的自动化。但由于智慧校园建设初期没有统一的数据标准,各个子系统在开发方式、技术及架构等方面各不相同,数据结构与类型极为复杂,给高校数据治理和管理带来了极大的困难,“信息孤岛”现象日益严重,高校各个部门之间的数据存在数据重复、数据源不用、数据冲突等问题。在这样的情况下,高校将多源、异构、分散的数据进行抽取、转换和融合,开始建立高校数据中心,为高校进行数据查询、分析、挖掘提供基础数据。将高校各个部门的数据进行有效融合,可以在一定程度上打破高校各个部门业务系统之间的“数据壁垒”,实现高校数据的实时共享,在高校数据中心建设的实际过程中,数据融合还存在很多问题,融合不仅涉及数据标准规范,还涉及数据动态同步来保障数据的实时共享。因此,解决目前高校数据共享及数据融合的难度成为目前智慧校园建设的重点。

2 相关研究和技术介绍

2.1 相关研究

数据融合的研究主要集中在数据标准和数据同步技术两个方面。李春子^[1]认为数据标准在高校数据管理中十分重要,并列举了数据标准实施的案例,提出了数据标准的制定策略,极大地推动了高校数据共享的步伐。张聪^[2]认为高校在智慧校园建设初期缺乏规划,依据国内外数据标准制定的经验,阐述高校数据标准的落地方案,提出高校数据标准制定和落地的实例。李华等人^[3]认为数据标准化建设是解决高校智慧校园建设中“信息孤岛”的唯一途径,从标准建设思路及路线和数据标准体系编制过程出发,提出了数据标准建设的时间方案。刘仲会^[4]从功能需求、体系结构与开发环境、主要技术及系统等方面调研并分析设计了基于 Windows Service 的高校管理软件。屠增辉等人^[5]通过对高校的数据进行分析,针对十分有价值的采取“开放共享,主动同步”的方式,运用 DataX 技术实现数据同步。贾卫忠等人^[6]认为 FTP 作为一种简单的数据同步技术,可以有效地解决高校信息化过程中出现的“信息孤岛”问题。魏李方^[7]根据不同场景,提出了触发器数据同步、定时任务、数据同步工具及定制数据同步程序四种常见的数据同步技术,并提出了不同场景的应用方案。目前关于数据融合的研究主要集中在理论研究与实践研究两个方面。在理论研究上还比较薄

弱, 没有形成完整的理论体系; 在实践研究上主要涉及不同领域数据融合的研究, 特别是物理传感器研究方面; 高校数据融合的研究主要集中在数据中心建设和智慧校园建设中的数据融合技术应用^[8]。

2.2 相关技术

数据融合是指将不同级别、不同来源、不同层次、不同结构的数据进行预处理, 最后进行关联及整合的过程^[9]。数据融合主要包括三点: (1) 数据融合的数据来源于不同的层次, 是多层次的数据融合; (2) 基于最终的数据融合目标进行数据的关联及整合; (3) 融合结束后要注意数据关联的准确率。数据融合相关的技术主要包括数据 ETL 技术和数据清洗技术^[10]。

数据清洗技术可以极大地提高数据的质

量, 是数据融合的基础, 数据清洗分为三层, 数据应用层、数据缓冲层及数据仓库的数据清洗。首先进行数据的校对及确定清洗流程, 接着验证数据清洗工作流程的合理性, 然后开始数据清洗, 最后将清洗后的数据装入数据仓库^[11]。

数据 ETL 技术是用来构建数据仓库的, 主要包含抽取、转换和加载三个步骤, 最终达到解决数据分散、标准不统一的目的^[12]。数据抽取是对数据源进行选择, 包括全量抽取和增量抽取; 数据转换是为了解决数据格式和编码不一致的问题并对数据进行清洗; 数据加载可以将清洗后的数据装载在数据仓库中, 实现数据真正的同步^[13]。数据 ETL 技术在应用过程中要注意抽取规则、清洗方法、转化方法 (见图 1)。

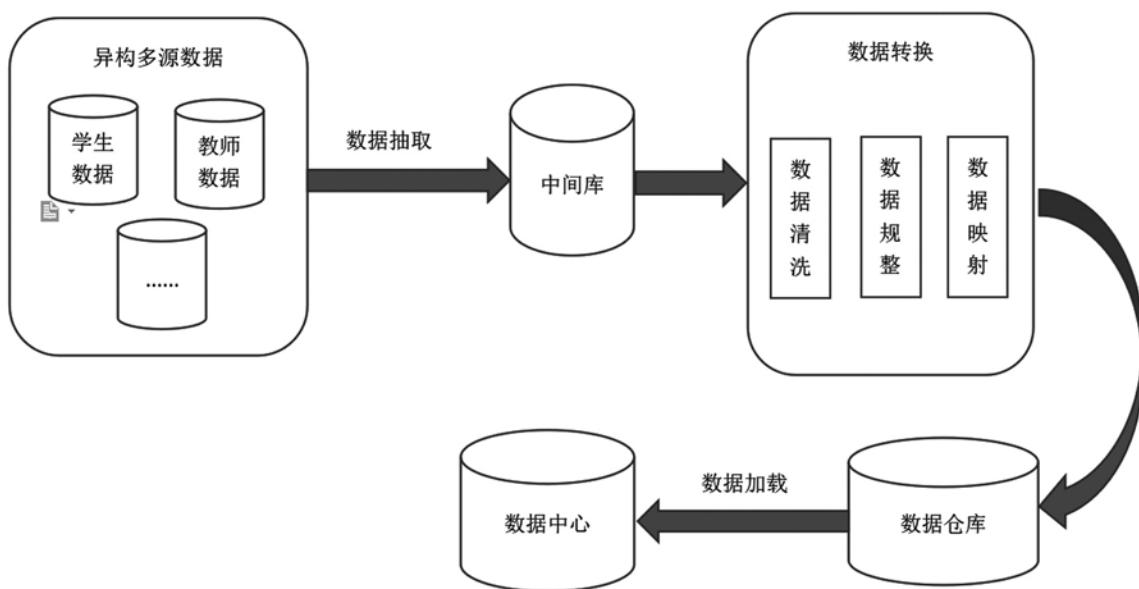


图 1 数据 ETL 技术路线图

3 高校数据融合现状与需求分析

3.1 高校数据特点和类型

3.1.1 高校数据特点

高校数据中心的建设推动了高校信息化建设的发展。高校各个部门的信息系统每天产生大量的数据,教学生活过程中也产生大量的数据。学生有姓名、学号、学习成绩、消费等数据;教师有科研、教学任务等数据,职能部门有财产资产管理数据、档案管理数据等^[14]。信息技术和大数据技术的发展推动高校每天产生大量的数据,主要呈现出以下几个特点。第一,数据存储量大,高校各部门业务系统的数据涵盖师生基本数据、师生教学数据、师生管理数据等多个方面。第二,数据类型多样,高校数据中心记载着教师的职工信息以及学生的各种信息,这些数据类型多样,包括学生的个人信息数据、学生的学习数据、学生的消费数据、教师的科研数据、教师的教学数据等,这些数据在不同的系统中的存储类型也不一样。第三,数据更新速度快,高校师生人数众多,在每天的教学生活中会产生大量的数据,数据更新速度极快。第四,数据结构复杂,信息技术的应用使高校各个部门产生的数据格式、类型日益丰富,造成数据结构复杂多样。第五,数据具有分析价值,高校数据中心存储着大量的师生数据,对这些师生的行为数据进行分析,不仅可以更好地为师生提供服务,还能推动高校信息化建设^[15]。

3.1.2 高校数据类型

高校数据来源丰富、数据量大,主要有基本数据、关联数据和统计应用数据。基本数据如教师个人信息、学生个人信息等,不会随时间发生变化,又称“静态数据”,其服务场景是指查询基本信息服务的场景,其服务对象主

要是高校师生及其相关的职能部门^[16]。关联数据如教师科研项目情况及学生考核情况等,统计应用数据如教师工作绩效评估及学生学习情况评估等,是经过数据分析后得到的数据结果,数据结构发生了变化,又称“结果数据”,其服务场景是基于两种信息,并经过统计和计算得到新的数据,提供基于数据的决策分析服务的场景^[17]。

3.2 高校数据融合的现状

高校数据中心的建设是高校信息化建设必要的一环,而数据融合是高效数据中心建设的关键。高校数据具有数据存储量大、数据类型多样、数据更新速度快、数据真实可靠、数据结构复杂、具有分析价值等特点。为了加快高校信息建设的步伐,高校各部门信息系统的数据在建设时缺乏统一的数据规划,高校“信息孤岛”“数据冲突”现象日益严重。高校数据融合可以为数据的存储管理提供一定的支持,高校开始利用数据融合技术建成高校数据共享中心,但是高校数据融合在实施过程中会遇到较多的问题和困难。

3.2.1 数据标准问题

高校各个部门数据标准问题主要体现在三个方面。第一,数据标准不同。高校各个部门引进信息系统的时间各有差异,各个部门之间的业务流程也相差较大,各个部门均是根据自己的需求建立标准的,部门与部门之间缺乏统一的数据标准,例如在A部门年龄是采用具体数值的形式,而B部门人员年龄则是采用年龄段的形式。导致各个部门之间不能很好地实现业务互通与数据共享。第二,缺乏制定数据标准的意识。高校大部分部门使用的信息系统都是由第三方公司开发的,造成数据标准不一致的现象。首先,各个部门找的第三方公司各异导致数据库的数据标准不同;然后,第三方公

司对学校各个部门的业务不熟悉,数据标准的制定也是根据经验,缺乏针对性;最后,高校在引入信息系统的初期并没有制定数据标准的意识。第三,现有的数据标准不够规范。随着高校数据共享中心的建设以及各部门数据共享需求的出现,高校开始制定一些数据标准,制定初期大都是借鉴国家标准和其他行业标准,因此标准缺乏科学性与前瞻性,数据标准不够规范,有待进一步完善。

3.2.2 数据冲突问题

高校数据共享过程中经常会出现数据冲突问题,导致数据共享的质量不高。数据冲突问题主要表现在三个方面。第一,错误值冲突。这类数据冲突指的是部分数据丧失了准确性,高校各部门很多数据需要手工录入或数据导入,在这个过程中极易出现错误,如性别、年龄等数据错误。第二,规范值冲突。这类数据冲突指的是同一个数据项出现不同的值,虽然不会影响到数据的质量,但是会影响数据的应用,例如性别要求写“男”或“女”,有的写的是“man”或“woman”,手工录入或原始数据导入均可能出现此类情况。第三,格式冲突。这类数据冲突大多也是手工录入或者数据导入出现的数据冲突现象,如出生日期、出生地等数据录入,格式不一样在数据共享的时候也会影响其效率。

3.2.3 数据同步问题

高校各个部门信息系统中的数据大部分都是属于关联数据和统计应用数据,数据变化较大,各个部门在数据共享的时候,经常会出现数据不同步的问题。例如本部门信息系统中的数据是实时更新的,但是并没有和高校数据中心的数据实现同步,中间存在一定的时间差,其他部门在使用这个数据的时候就会出现数据不一致的情况。

3.2.4 数据人员问题

高校从事数据工作的人员大都集中在网络与信息中心,其他各个部门缺乏专业的数据工作人员,一般由部门内部的某个人代替数据类工作,这为数据融合带来了较大的困难。一方面,从事数据工作的专业人员的数量有限,各部门的信息技术人员一般负责信息系统的维护,业务人员主要是利用信息系统完成日常工作,没有多余的人员制定部门的数据标准。另一方面,缺乏专业的数据分析人才,学校在数据融合的过程中需要考虑部门岗位设置的合理性,因此,没有引入大量的专业数据分析人才。

4 基于数据共享的高校数据融合策略研究

4.1 构建数据融合整体架构

高校业务数据库较多,各部门均建有自己的信息系统和数据库。常见的业务数据库有人事数据库、教务数据库、财务数据库、科研数据库、图书馆数据库等,这些数据是源数据,一定要确定源数据的准确与权威性。数据存储主要是设计数据湖作为中间库,中间库存放的是高校各部门业务数据库的原始数据与数据中心更新后的数据,中间库的设计可以很好地避免数据在融合的过程中各个部门的业务信息系统同时获取数据时,对高校各个部门的业务信息系统的性能产生影响,保障数据安全。数据融合主要是通过数据抽取、数据清洗、数据转换步骤,最终将数据传输到数据湖中,形成标准化的数据,方便后面数据的应用。数据应用主要包括数据的查询、分析、挖掘,可以为高校管理人员提供科学化决策的依据。因此,基于数据共享的高校数据融合的全过程主要包括三步:第一,抽取高校各业务系统的原始数据;第二,原始数据同步到数据湖中;第三,

数据湖中的数据同步到数据中心。

建立基于数据共享的高校数据融合整体架构，可以很好地为高校决策和全校师生服务。整体架构主要包括了高校业务数据库、数据存储、数据融合与数据应用（见图2）。首先，参考前期制定的数据标准规范，对高校数据中

心的数据统一建模，设计好各数据库之间的关系。其次，将高校各个部门的业务员数据库中的数据同步到中间库（数据湖）。最后，通过数据抽取、数据转换、数据加载获取数据湖中的数据并上传至数据中心，进行查询、统计分析、挖掘等数据应用。

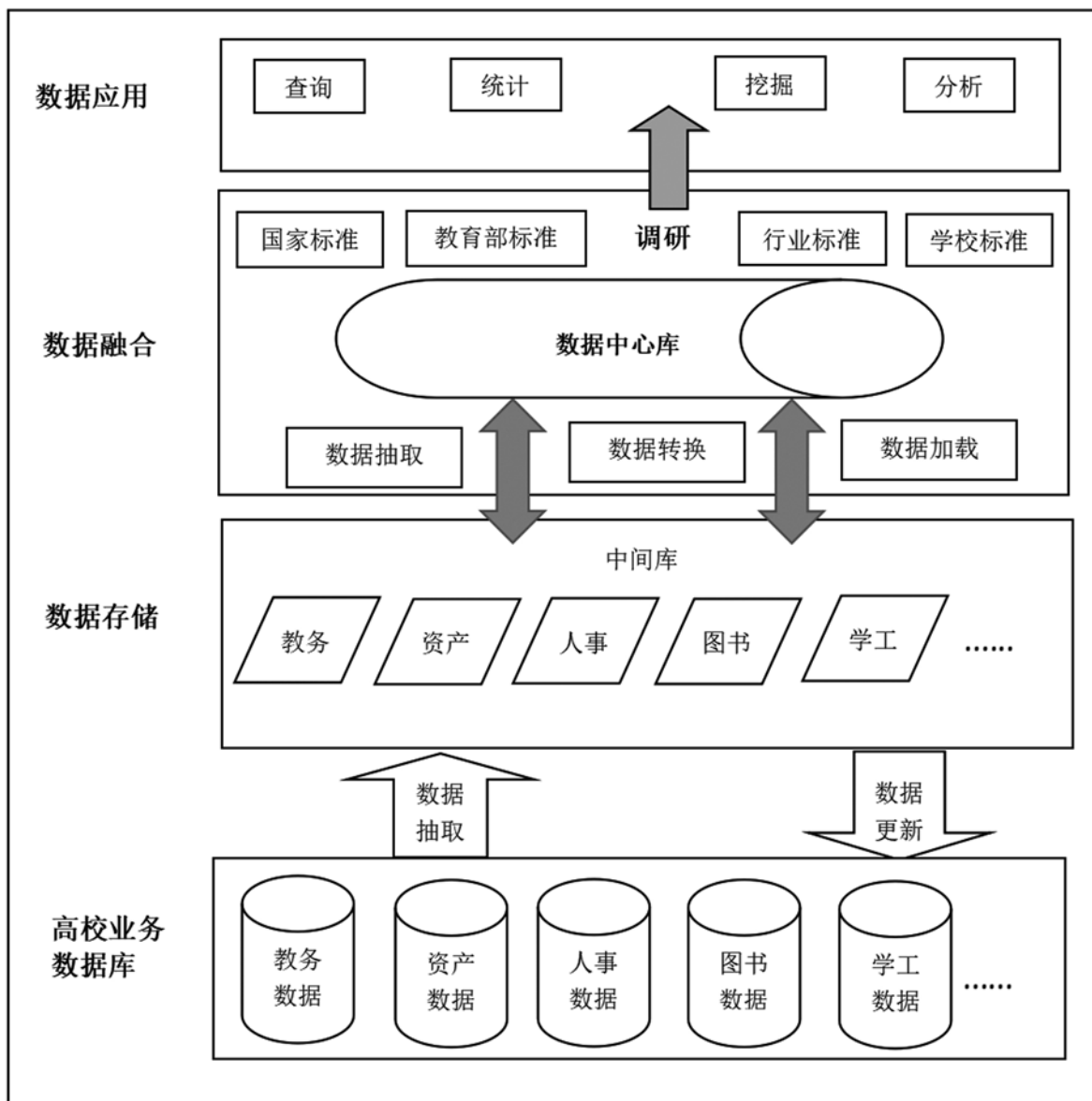


图2 基于数据共享的高校数据融合整体架构框架

4.2 制定数据标准规范

制定数据标准是构建高校数据中心的前提,也是高校数据融合的第一步,是数据同步的基础。高校应该从部门需求和学校发展出发,统计现在与将来的各类数据信息,参考国家、行业、教育部门以及学校的标准,制定基于学校实际情况的数据标准,为数据融合奠定基础。

一方面,确定数据标准设计的原则。对高校学习、业务、资源等数据制定一套比较完整的数据标准体系,采用自顶向下的思路可以将数据分为数据集、数据类、数据子类、数据项四个层次。数据标准体系应用到高校各个业务数据库中,在其中担任数据字典的作用,数据标准的建设应该满足兼容性、唯一性、可扩展性、规范性原则。兼容性原则要求高校各个部门的业务数据库的数据标准要与国家、行业、教育部等颁布的标准相兼容。唯一性原则要求高校各部门业务数据库中数据项的编码对象必须有唯一的代码。可扩展性原则要求高校各部门业务数据中的数据标准应该遵循数据动态变化的特点,内容可以更新和扩充。规范性原则要求高校各部门业务数据库数据标准中的编码体系、类型、格式等应该规范一致。

另一方面,建立规范的数据标准体系。基于数据标准的设计原则,将高校涉及的元数据进行划分,形成各类数据标准集合,并针对每一个数据集设计唯一的代码,再为数据集下面的数据类型设计编号及代码,最后对每个数据类型下面的数据项设计编号及字段代码。高校数据标准集一般包括高校概况数据集、学生管理数据集、教职工管理数据集、教学科研管理数据集、财务管理数据集、资产管理数据集、图书管理数据集等。以学生管理数据集为例,为其设计唯一的编号“GXXSGL”,其数据集下面的数据类型——学生基本数据类型、学籍

数据类型、学位学历数据类型、实践活动数据类型等分别编号为“GXXSGL01”“GXXSGL02”“GXXSGL03”“GXXSGL04”。学生基本数据类型下面的数据项——姓名、学号、性别等分别标号为“XM”“XH”“XB”。

4.3 制定合适的数据融合规则

数据冲突为数据共享和数据融合带来了极大的挑战,不同数据冲突类型对应不同的解决方案。通过数据清洗技术可以有效地识别错误值冲突,可以删除错误值并将这个值赋为空值,然后补充空值即可。对于不规范值冲突,可以通过相关的语义词典及相应的规则来解读。通过数据清洗技术可以解决格式规范的问题,将高校各个部门的数据定义为统一的格式。由此可以看出,制定恰当的数据融合规则可以有效地解决数据冲突的问题。首先,制定基于数据格式的数据融合规则,在数据清洗阶段,通过定义统一的数据项格式,如出生日期、入学时间等规范的数据项,根据数据所需要的实际格式和适用场景来定义标准的数据格式。然后,制定基于语义的数据融合规则。在数据初次融合的阶段,将语义比较相同的数据值认为是同一个值,然后在数据字典中不断补充和更新。最后,制定基于权威数据源的数据融合规则。在数据再次融合的阶段,如果出现部分数据失真的情况,可以通过数据源及相关部门与数据相关的工作人员的经验来确定权威性,来判定数据的准确值。

4.4 科学采用数据同步技术

数据同步技术是实现数据共享和数据融合的技术保障。高校的关联数据和统计应用数据属于不断变换的数据,需要采用数据捕获方法捕获数据。根据高校各个部门业务数据库中数据的开放程度与类型,选择常见的时间戳法、日志法或控制表来捕获变化的数据,捕获后的

数据就可以实现数据同步。高校数据同步的方法主要包括全量数据同步和增量数据同步。根据基于数据共享的高校数据融合架构可知,在数据融合的过程中有业务系统数据同步到中间库和中间库的数据同步到数据中心两次数据同步。

采用数据同步技术将业务系统数据同步到中间数据库。将高校各部门业务系统数据库中的数据同步到中间库中时,对数据的实时性要求不高,为了保障数据的安全,采用定时增量同步的方式实现同步。这个过程主要包括两个步骤:第一,捕获变化数据模块,在源数据库中建立一个触发器,利用控制表法去捕获变化数据;第二,同步数据载入模块,编写 SQL 代码读取高校各部门业务数据库中的数据信息,实现中间库与高校各部门业务数据库的同步。

采用数据同步技术将中间库与数据中心的数据同步。高校数据中心存储的数据已经具有一定的标准规范,因此,数据同步采用全量同步的方式。这个过程主要包括数据关联、数据处理和数据载入三个步骤:第一,数据关联,关联数据主要是指数据项中的数据记录的关联,如果数据发生变化,通过其他表中的数据进行连接更新;第二,数据处理,关联数据同步到高校数据中心后的数据应该是标准规范的数据,利用数据筛选、转换等操作,形成标准规范的数据集;第三,数据载入,将更新后的数据和规范的数据载入目标数据库中,实现中间库到数据中心的数据同步。

4.5 培养数据专业人才

在高校智慧校园建设的过程中缺乏数据融合专业人才的参与,学校通过培养数据专业人才可以很好地解决数据融合过程中数据效率和质量的问题。首先,学校可以在各个部门引入

或者培养专门的数据管理与分析人才,高校数据共享与融合离不开专业的数据管理人员的支持,需要专业的人员对数据进行实时的管理与维护。例如可以在教务部门培养专门的数据分析人员,可以利用数据分析技术和信息技术对学生的成绩、教师的课程安排进行分析,及时发现教学中的不足,可以给教学工作提供科学的建议。第二,建立专门的数据管理与分析部门,分别从宏观和微观上对校园的数据进行分析,还可以建立专门的校情分析信息系统或数据挖掘信息系统,及时发现学校教学、生活、工作中的不足,为全校师生提供更加全面便捷的服务。此外,培养数据专业人才是数据融合过程中数据质量的保障,数据管理质量的高低会影响到数据融合的结果,数据管理对计算机能力有一定的要求。因此,在培养数据专业人才的过程中,不仅要重视数据素养的培养,还要注意计算机素养的培养。

5 结语

加快高校智慧校园的建设就必须实现数据融合,大数据时代发展的背景下,高校大数据具有一定的价值,分析和挖掘高校数据,以便更好地为师生提供个性化的服务是智慧校园建设的重点。本文从数据融合中的数据标准、数据冲突、数据同步与数据工作人员现状出发,分析其存在的问题,并结合实际提出基于数据共享的高校数据融合路径。一方面,高校需要基于学校实际制定科学的数据标准规范,确定各部门数据集合、数据类和数据项的标准。另一方面,采用数据同步技术并培养专业的数据工作人员来提高数据的质量。数据融合是智慧校园建设到智能校园建设的基础,高校数据融合与数据共享的实现是一个长期的工作,目前大部分高校还处于探索起步阶段,需要经过不

断地优化。制定规范的数据标准与应用数据同步技术只能阶段性地解决高校数据共享中“信息孤岛”的问题,后续还要将研究着眼于一个高校数据的全民融合乃至区域高校数据的共建共享,将信息技术、大数据、云计算的技术运用到高校数据融合之中,帮助学校管理人员科学地决策,为师生提供个性化的数据服务,推动智慧校园的建设步伐。

参考文献

- [1] 李春子. 适用于高校的数据标准制定与实现[J]. 信息技术与标准化, 2010(5): 58-60.
- [2] 张聪. 高校数据标准的制定与落地[J]. 中国教育信息化, 2021(15): 48-51.
- [3] 李华, 刘波, 龚艺, 等. 高校教育管理数据标准建设实践[J]. 安徽电子信息职业技术学院学报, 2017, 16(2): 97-102.
- [4] 刘仲会. 基于Windows Service的高校管理软件间数据同步系统的设计与实现[J]. 天津职业院校联合学报, 2019, 21(9): 102-108.
- [5] 屠增辉, 刘淼, 陈小丽. 基于DataX的异构教学数据同步研究和应用[J]. 中国现代教育装备, 2021(19): 114-118.
- [6] 贾卫忠, 李存华. 一种基于FTP的数据同步技术[J]. 计算机时代, 2008(11): 57-59.
- [7] 魏李方. 数据同步技术在高校数据中心的应用[J]. 电脑知识与技术, 2020, 16(10): 266-268.
- [8] 华俊. 超融合架构技术在高校数据中心建设中的创新应用[J]. 电脑知识与技术, 2021, 17(34): 30-31.
- [9] 杨东. 应用型高校教学管理信息化体系数据互通融合的实践探索[J]. 长春教育学院学报, 2021, 37(11): 32-37.
- [10] 黄小明. 基于校园数据治理下融合门户的建设方案——以宁波工程学院为例[J]. 信息技术与信息化, 2021(11): 52-55.
- [11] 武全全, 唐双龙. 高校语文学科融合在大数据背景下的教学改革研究[J]. 食品研究与开发, 2021, 42(22): 251.
- [12] 杨彩云, 方建文, 周小龙, 等. 产教融合视角下的工程实践教育质量提升——以地方应用型高校的数据科学与大数据技术专业为例[J]. 高教学刊, 2021, 7(31): 69-72.
- [13] 户利利, 张聪. 超融合技术在高校数据中心的应用[J]. 微型电脑应用, 2021, 37(7): 186-188.
- [14] 银黔丰. 高校数据融合中数据标准与数据同步技术研究[D]. 长沙: 湖南师范大学, 2021.
- [15] 史蕊. 超融合云计算在高校数据中心的应用与实践[J]. 软件, 2021, 42(4): 42-44.
- [16] 蒲飞, 涂旭东, 陈苗, 等. “七步法”助力高校走出数据融合困境实践——以重庆医药高等专科学校为例[J]. 中国教育信息化, 2021(5): 46-49.
- [17] 吴云峰. 多维数据融合的高校科研数据管理模型构建与运行保障研究[J]. 情报科学, 2020, 38(12): 110-115.