文章编号：2095-2163(2020)05-0218-06 中图分类号：TP315 文献标志码：A

高校数据仓库多维数据建模分析

张 军，王芬芬
（湖南铁道职业技术学院 图书信息中心，湖南 株洲 412001）

摘 要：为满足高校对各业务系统所有产生的操作型数据进行集成、统计和分析需求，需应用数据仓库技术，完成业务系统数据的采集与治理工作，为上层应用提供规范、有效的数据服务。基于此，文章结合高校业务数据的特点和实际应用需求，给出了高校数据仓库平台的架构设计，针对高校数据分析需求多变、统计复杂的特点，文章设计了集中抽取、分区治理、多维建模的建设方案，着重分析了主题数据区的维度模型构建方法，设计了教师基础数据星型模型和学生学籍数据雪花模型。最后给出了数据清洗策略，为高校数据仓库的建设提供了参考和借鉴。

关键词：数据仓库；数据分析；维度模型；雪花模型

Analysis of multidimensional data modeling in university data warehouse

ZHANG Jun, WANG Fenfen
(Information and Technology Center, Hunan Railway Professional Technology College, Zhuzhou Hunan 412001, China)

【Abstract】 In order to meet the needs of integration, statistics and analysis of all the operational data generated by various business systems in Colleges and universities, it is necessary to apply data warehouse technology to complete the collection and management of business system data and provide standardized and effective data services for the upper application. Based on this, combined with the characteristics and practical application requirements of university business data, this paper presents the architecture design of university data warehouse platform. According to the characteristics of the changeable and complex data analysis needs of universities, this paper designs a construction scheme of centralized extraction, partition governance and multidimensional modeling, and focuses on the analysis of the construction method of the dimensional model of the subject data area, the star model of teachers´ basic data and the snowflake model of students´ status data are designed. At last, The data cleaning strategy is given, which provides reference and reference for the construction of data warehouse in Colleges and universities.

【Key words】 Data Warehouse; Data Analysis; Dimension Model; Snowflake Model

0 引 言

随着高校各类信息系统的深入使用，已经累积了大量的数据，有效组织与分析这些数据是当前高校信息化建设的主要任务。很多高校前期已经建立了不同程度的，用于数据交换与共享的数据中心平台，该平台能够实现简单的数据集成，但当前高校业务数据已经呈现出历史数据量大、数据异构、数据冗余且不一致，在统计方面也存在数据统计维度多、统计路径多样化等特征。在这种条件下，要实现跨时间、跨业务的综合统计分析是一项十分困难的工作。构建高校数据仓库系统能有效解决上述问题，数据仓库的 ETL(Extract-Transform-Load)过程能有效的解决数据异构、冗余、不一致等问题，同时数据仓库能够在各种粒度上为多维数据的交叉分析提供支持[1]，并且所积累的大量历史数据能够为数据挖掘提供完善的数据样本集。

1 架构设计

数据仓库主要面向统计分析和数据挖掘，为高校的教学和管理决策提供支持。其数据来源为高校内其它操作型业务数据库和数据文件，这些数据按照高校制定的数据标准，经过清洗、转换加载至数据仓库中，为上层应用提供支撑。本文所构建的高校数据仓库主要包括源数据层、数据仓库层和数据应用层，如图 1 所示。

源数据层。该层为各业务系统的数据库，是数据仓库层的数量来源。

数据仓库层。该层主要分为近源数据、标准数据和主题数据 3 个区域。近源数据区贴近业务系统源数据，保存了各个业务系统的数据明细，与源业务系统的数据结构基本保持一致，唯一不同之处是在原有基础之上添加了时间戳，形成不同版本的历史数据。标准数据区是数据仓库的核心数据区，是按照单位制定的数据标准对近源数据进行标准化处理后的结果，该层数据符合数据库第三范式建模要求。主题数据区对应宏观的分析领域，通过对标准数据进行重新组织或汇总，为不同主题的数据建立维度

基金项目：2018 年度湖南省教育厅科学研究项目(18C1528)。
作者简介：张 军(1984-)，男，硕士，讲师，主要研究方向：数据挖掘、数据库技术。
收稿日期：2020-01-10

汇总数据区，以满足上层应用对数据的多样化需求，该层采用维度建模的方法构建。

图 1　高校数据仓库的架构设计

Fig. 1　The architecture design of university data warehouse

数据应用层。该层为用户和数据仓库层建立交互界面，依据用户请求访问仓库内的数据，生成各类数据统计报表，实现对数据多维度、多层次的分析和隐性知识的挖掘。包含了多维分析、统计报表、数据挖掘，以及为其它应用提供的数据接口[2]。

2　数据建模

数据建模是构建数据仓库的核心工作之一，通过数据建模，能够使高校建立全方位的数据视角，勾勒出高校各部门间的内在联系，同时能够有效解决各业务数据的一致性问题。另外，数据建模可以分离出底层技术的实现和上层业务的展现，能够有效应对业务的变动，提高数据仓库的灵活性。

依据图 1 所设计的高校数据仓库架构，在数据仓库层有 3 个不同的数据区域，分别为近源数据区、标准数据区和主题数据区，其中近源数据区和标准数据区主要是完成数据的抽取与转换，对数据进行标准化处理，其建模方法基本是采用传统的数据库范式建模法。灵活多变的分析需求是主题数据区建模所必须应对的问题，依据高校具体的业务数据特点以及分析需求划分主题域，每个主题对应 1 个宏观的分析领域，在主题数据区中为主题建立其所需的事实表与维度表，确定关联关系，建立多维数据模型，进而为上层应用提供数据服务，其一般过程如图 2 所示。

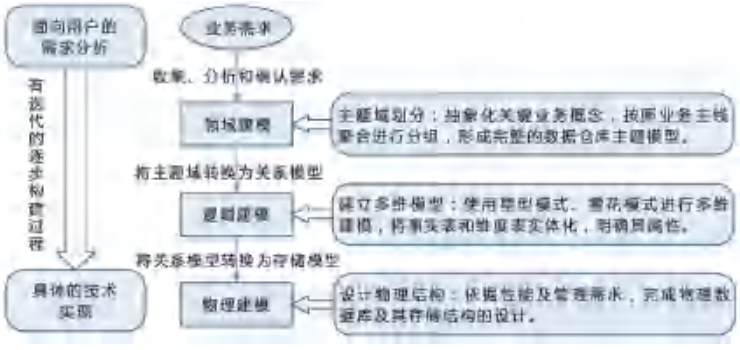

图 2　主题数据区建模过程

Fig. 2　Modeling process of subject data area

2.1　领域建模

领域建模就是对业务系统充分了解后，结合高校数据仓库的建设需求，对关键业务抽象化，按照业务主线聚合进行分组，将业务数据进行综合、归类，最终形成面向相应宏观分析领域的各个主题域。主题域划分通常采用树形结构，采用逐级细分的思路进行设计。本文依据高校的核心业务，定义了 1 个公共主题和 5 个业务主题，如图 3 所示。公共主题包括了学校的基础数据和标准代码集，这些标准代码集参考了国标和教育部相关标准，也包含了学校自定义的校标[3]。部分公共维度也包含在代码集中，比如时间、地理位置、专业、学历学位等公共维度，它们通常与多个主题事实表产生关联，形成多维分析模型。

2.2　逻辑建模

逻辑模型是在充分理解分析主题与用户需求的基础上，确定分析粒度，为每个主题事实选取分析维度，设计事实表及其相关联的维度表。事实就是对分析主题的度量，其度量属性的值就是进行分析处理的对象，事实表的设计以能够正确记录历史信息为准则。维度则是对分析主题所属类型的描述，是分析者观察事实的角度，维度表的设计是以能够用合适的角度来聚合事实内容为准则[4]。事实表和维度表的设计是逻辑建模的关键，其设计的好坏也直接影响到整个数据仓库系统的性能以及数据分析效果。

图 3 高校数据仓库主题域划分

Fig. 3 Subject domain division of university data warehouse

多维数据模型依据事实表和维度表不同的组织形式,通常有三种设计模式,星型模式、雪花模式和事实星座模式。本文以高校业务中比较核心的数据分析主题来阐述三种不同模式下的多维数据模型建立。

(1) 星型模式。星型模式的基本结构就是事实表位于中心,维度表围绕在事实表周围[5],这种模式能够直观的展示数据的多维功能。教师主题下的教师基础数据能够从多个维度对高校师资结构进行统计分析,同时也可以作为教学、科研等与教师相关主题的教师维度。教师基础数据涉及的维度较多,如学历、学位、学科、职称、岗位、民族、籍贯等等,适合采用结构简单的星型模式进行建模,不仅可以直观的反映出业务逻辑,还便于对不同的维度进行灵活的组合分析。为跟踪教师基础数据的变化过程,记录和保留活动数据的历史信息,本文事实表的设计引入了缓慢变化维的方法来捕获变化数据。在事实表中加入开始时间、结束时间和版本 3 个属性来实现记录维度的缓慢变化,其中开始时间和结束时间标记了活动数据在该时间段内处于某一状态,版本记录了活动数据经历的历史状态顺序[6]。教师基础数据的星型模型如图 4 所示。

图 4 教师基础数据星型模型

Fig. 4 Star model of teachers´ basic data

（2）雪花模式。雪花模式和星型模式有类似的逻辑模型，也是由事实表和维度表组成。在雪花模式中，维度表中低基数的属性被移除，形成单独的表，基数是指表中一个属性不同值的个数，这项操作就是维度规范化。当维度被规范化成多个关联的表，即形成了以事实表为中心的雪花型结构。维度规范化将维度表中重复的组分离成一个新表，这些通过分解形成的表连接到主维度表而不是事实表，有效的减少了数据冗余，但却不可避免的增加了表的数量，在执行查询时，不得不连接更多的表。但是规范化减少了存储数据的空间需求，提高了数据更新的效率。

以学生学籍事实表（FACT _ STUDENT _ STATUS）及其教学班级维度表（DIM_CLASS）为例，进行雪花建模，阐述维度规范化的过程，分析其在存储空间及更新效率上的优势。以某高校为例，该校有在校生 20000 人，12 个二级学院，25 个教学系部，共计 400 个教学班级。如果以星型模式进行建模，事实表有 20000 条记录，教学班级维度有 400 条记录，共计 20400 条记录，每个学生所属的二级学院以及教学系部作为教学班级的属性，显式的存放在教学班级维度表中。对教学班级维度进行规范化处理，建立二级学院维度表（DIM_COLLEGE）和教学系部维度表（DIM_COLL_DEPART），事实表没有变化，总的记录数变为 20437（20000+400+12+25）条记录，规范化增加了新表，总的记录数也增加了，但是不难看出，在教学班级维度表中存放的不再是二级学院和教学系部具体的属性信息，而是它们的主键值，具体的属性信息统一存放在其相关的维度表中，这样就大大减少了数据存储所占用的空间，教学班级的数量越大，这种空间优势就越明显。在数据更新方面，如果学校发生了院系调整，只需更新二级学院及教学系部维度表，对数据量较大的事实表的影响是十分微小的。

实际上，星型模式是雪花模式的一个特例（维度没有多个层级）。雪花模型的主要缺点是维度属性规范化增加了查询的连接操作和复杂度。相对于平面化的单表维度，多表连接的查询性能会有所下降。但雪花模型的查询性能问题近年来随着数据浏览工具的不断优化而得到缓解。学生学籍数据的雪花模型，如图 5 所示。

图 5　学生学籍数据雪花模型

Fig. 5　Snowflake model of student status data

（3）事实星座模式。高校数据仓库由多个主题构成，包含了多个事实表，很多事实表里包含了大量公共的维度，这些维度供多个事实表共享使用，形成了多个星型模式的汇集，这种结构就是事实星座模式，也称为星系模式。以高校财务明细事实表和科研项目事实表为例，它们之间存在着大量的公共维度，比如项目负责人、项目类别、项目来源、项目级别、立项时间、结项时间等等，多个事实表与多个公

共维度交叉连接,是数据仓库构建过程中常用的建模方式。

多维数据模型是数据仓库的核心,也是 OLAP(联机分析处理)的灵魂。上述三种多维数据的建模方法都是由一组维度和事实的集合组成的,该模型可以用一个 $n(n \geqslant 2)$ 维数据立方体表示,数据立方体中的维来自维度表,空间中的点来自事实表,每个点(以取 $n = 3$ 为例,1 * 1 * 1)包含事实数据,称为存储单元。多维数据分析的核心操作就是对数据立方体进行钻取(Drill-down)、上卷(Roll-up)、切片(Slice)、切块(Dice)以及旋转(Pivot)。钻取和上卷是通过改变维度的层次,调整分析粒度来观察数据,切片是通过固定数据立方体上某一维度上的选定值,观察数据在剩余维度上的分布情况,如果是对两个及以上的维度执行选择就是切块操作,旋转即是变换维度在数据立方体上的方向[7]。

针对高校业务数据结构繁杂,且存在着大量聚合分析的特点,本文在图 5 的学生学籍数据雪花模型中进行了维度层次结构的设计。以教学班级维度为例,教学班级隶属于教学系部,教学系部隶属于二级学院,层次之间通过外键连接,能够实现“教学班级→教学系部→二级学院”的上卷或“二级学院→教学系部→教学班级”的钻取。建立维度层次关系,可以用来定义切片路径,数据立方体可以通过教学班级、教学系部或者二级学院进行切片分析。

2.3 物理建模

物理建模的主要工作是将所设计的逻辑模型在合适的数据库管理工具中实现,包括选择合适的数据库管理工具,设计数据表的结构及其属性类型,建立用于快速访问的索引策略,明确数据的存储方式及存储位置,制定实施数据的装载与清洗策略。

依据本文在数据仓库层中所设计的分区域存储和治理数据的策略,需要创建的物理表主要包括如下几类:

(1)为满足数据仓库元数据管理和数据 ETL 的需求,所创建的配置表、日志表等。

(2) 在近源数据区用于存储原始数据的源数据表。

(3) 在标准数据区用于存储对源数据表进行清洗和转换的标准数据表。

(4) 在主题数据区用于存储多维数据模型所生成的维度表和事实表。

在具体设计过程中,需要对物理模型中的数据定义和数据格式进行规范化处理,也包括所遇到的一些设计共性问题,如在物理模型中所需要的主键是采用自然键还是代理键。自然键就是用实体现有的属性组成键值,在业务概念上是唯一的。代理键就是新增一列不具有业务含义的键值表示数据唯一。本文设计的大多事实表都是采用自增序列的代理键为主键,因为高校业务繁多,业务需求变更频繁,代理键不与业务产生耦合,业务需求的变更对其不会产生影响,更容易维护。另外,时间戳字段也是一个设计共性问题,数据的变化一般是发生在字段一级的,如果给每一个字段盖上一个时间戳,虽然能够最详细的记录标识数据的变化,但会大大增加数据的存储量,采用在行一级上添加时间戳,当数据发送变化时,时间戳字段同步更新,通过系统时间与时间戳字段的值来决定所抽取数据。

在索引创建策略上,按照索引使用的频率,由高到低逐步添加,使用主关键字和外部关键字建立索引,根据实际情况可以设计多种索引结构。在数据的具体存放位置上,将索引和数据表分开存放,索引存放在高速存储设备上,数据表可存放于一般存储设备,以加快数据的查询速度。

3 实现与应用

在上述数据模型的基础上,构建高校数据仓库还需要完成数据 ETL 的实施,ETL 是将各个业务中的异构数据源经过抽取、转换、加载到数据仓库的过程。ETL 是数据仓库实施的核心内容,常用的开发工具有 Oracle 公司的 ODI(Oracle Data Integrator)、开源工具 Kettle 等,也可以直接编写存储过程。本文选择 ODI 作为主要的 ETL 实现工具,但对逻辑复杂,且对执行效率有较高需求的 ETL,则直接使用存储过程来完成。

完成从各业务系统中抽取源数据后,对数据进行清洗和标准化也是一项重要的工作。数据清洗主要对源数据中出现的残缺数据、错误数据、重复数据以及违反逻辑规定的数据等问题数据进行统一的处理[8]。表 1 给出了针对高校业务系统常见的数据问题,以及对其所采取的清洗策略。数据标准化就是依据制定的信息标准对清洗后的数据进行规范化处理,如不同业务系统的同一数据的数据格式或使用的数据字典可能不一致,就需要将其按照数据仓库的信息标准进行规范化处理。

完成数据仓库各个层次的数据处理后,就可以为上层应用提供数据服务了,主要包括一些数据查询系统、在线分析系统、决策支持系统、数据挖掘与数据接口等。

表 1　数据清洗的常见问题及策略

Tab. 1　Common problems and strategies of data cleaning

问题类别	主要表现	清洗策略
残缺数据	业务系统未做非空约束，导致很多空值字段出现。	(1)由业务系统使用部门组织补全录入； (2)在该字段所对应的信息标准代码表中新增一个代码字段，将该字段设置一个统一的代码值； (3)直接将该字段予以删除。
错误数据	业务系统数据字典不规范或未使用选择录入导致的超出字典表范围。 业务系统未做录入检查，导致数据格式错误或数据越界。	(1) 要求业务系统规范字典表并对错误数据进行修正； (2) 导出错误数据至 Excel 文件交由业务部门修正。
重复数据	业务系统中未建立有效的主关键字	(1) 要求业务系统建立数据完整性约束； (2) 导出重复数据由业务部门确认。

4　结束语

本文基于典型的数据仓库构建技术，结合高校具体的数据统计与分析需求，针对高校业务系统零散、数据类别繁杂的特点，提出将数据仓库分为近源数据区、标准数据区和主题数据区 3 个区域，每个区域的数据具有不同的特点，同时采取不同的治理策略。多维数据建模是构建数据仓库核心内容，它能够满足对数据进行多层次、多角度的分析需求，本文选取了高校教师基础数据和学生学籍数据作为建模分析对象，分别给出星型模式和雪花模式的多维数据模型，由于篇幅有限，所构建的数据模型只列出了主要的关键字段，但它依然可以作为高校在构建数据仓库时进行数据建模的参考。本文的数据仓库架构以及所使用的多维数据建模方法已经应用于某高校的大数据分析平台，能够灵活的分析与统计高校业务数据，自动生成各类复杂的数据报表。

参考文献

[1] 王珊珊，孙其伟，陈云. 高校数据仓库构建与应用研究[J]. 华东师范大学学报(自然科学版)，2015，3(S1)：509-515.

[2] 赵宏斌，白开峰，崔丙锋，等. 基于 DAMT 的企业级数据仓库建设关键路径研究[J]. 江西师范大学学报(自然科学版)，2018，42(6)：634-638.

[3] 张端鸿，刘波，卞月妍. 院校数据仓库架构与建设的过程研究[J]. 高校教育管理，2017，11(2)：26-33.

[4] 修国林，黄雨笋，李国清，等. 基于敏捷型 BI 的矿业集团生产信息分析模型[J]. 中国矿业，2017，26(10)：30-37.

[5] 黄鹏飞. 复杂信息系统的数据提取，建模及应用[D]. 杭州：浙江理工大学，2018.

[6] 洪文波. 基于缓慢变化维的 BI 数据仓库建模平台设计与实现[D]. 北京：北京工业大学，2017.

[7] 唐秀忠，陈洪磊，陆玉发. 基于 OLAP 的高校数据分析与决策支持系统研究[J]. 现代电子技术，2019，42(2)：155-158.

[8] 邓嘉明，叶忠文，王荣华. 以数据聚合为核心的高校智慧校园体系建设[J]. 现代电子技术，2019，42(3)：134-138.

(上接第 217 页)

[5] 任志英，高诚辉，申丁，等. 双树复小波稳健滤波在工程表面粗糙度评定中的应用[J]. 光学精密工程，2014，22(7)：1820-1827.

[6] 王晓强，李艳娜，崔凤奎，等. 基于二代小波的表面粗糙度信息提取[J]. 河南科技大学学报(自然科学版)，2015，36(3)：14-17+5.

[7] 黄敏超，高美凤. 基于小波阈值组合滤波器的光谱去噪方法[J]. 江南大学学报(自然科学版)，2015，14(2)：136-140.

[8] 刘国宏，郭文明. 改进的中值滤波去噪算法应用分析[J]. 计算机工程与应用，2010，46(10)：187-189.

[9] 陈乃金，周鸣争，潘冬冬. 一种新的维纳滤波图像去高斯噪声算法[J]. 计算机系统应用，2010，19(3)：111-114.

[10] 胡松，江小炜，杨光，等. 滑动平均滤波在微弱脉冲信号检测中的应用[J]. 计算机与数字工程，2007，35(10)：169-171，193.

[11] 桑庆双，程健. SG 滤波在朗缪尔探针信号处理中的应用[J]. 计算机工程，2011，37(17)：220-222，226.

[12] 袁开明，舒乃秋，孙云莲，等. 基于阈值寻优法的小波去噪分析[J]. 武汉大学学报(工学版)，2015，48(1)：74-80.

[13] DONOHO D L，JOHNSTONE I M. Adapting to unknownsmoothness via wavelet shrinkage[J]. Journal of AmericanStat.Assoc.，1995，12(90)：1200-1224.

[14] 于笃发，邵建华，张晶如. 基于小波自适应阈值图像去噪方法的研究[J]. 计算机技术与发展，2013，23(8)：250-253.

[15] 郝建军，刘勇刚，廖刚，等. 一种改进小波阈值函数的信号去噪[J]. 重庆理工大学学报(社会科学版)，2019，33(4)：93-97.