数据仓库在高职院校中的应用①

杨仁怀, 郎川萍

(四川交通职业技术学院 计算机工程系, 成都 611130)

摘 要: 以四川交通职业技术学院为例, 主要讨论了高职院校信息化现状、建设数据仓库的基本策略、确定数据仓库的主题、数据仓库的粒度、选择数据仓库模型等. 基于四川交通职业技术学院信息化建设现状, 由数据集市开始, 采用“自底向上”的策略建设数据仓库, 是一种可行的方法. 最后使用 OLAP 对学生成绩和收费数据进行了分析.

关键词: 高职院校; 数据仓库; 联机分析; 建设

Application and Research of Data Warehouse Technology in Higher Vocational Colleges

YANG Ren-Huai, LANG Chuan-Ping

(Computer Engineering, Sichuan Vocational and Technical College of Communications, Chengdu 611130, China)

Abstract: Based on current college IT construction in Sichuan Vocational and Technical College of Communications, this paper discusses strategies, objectives, granularity and model of deploying a data warehouse. First it begins with Data Mart; this paper presents a good practice of building a data warehouse with the strategy of bottom to top. finally, it uses OLAP to analyze students' score and charge data.

Key words: vocational and technical college; data warehouse; OLAP; construction

“数据仓库”这一概念始于上世纪 80 年代中期, 被誉为 “数据仓库之父”的 W.H.Inmon 在 1990 年出版的《Building the Data Warehouse》一书中对数据仓库的定义是大家最为广泛接受的, 即数据仓库是 “面向主题的、集成的、随时间变化的、不容易丢失的数据集合, 支持管理部门的决策过程[1]. 近几年来, 数据仓库在许多行业得到了广泛应用, 例如零售、金融、制造、交通、电信、保险和医疗健康等行业. 但数据仓库在教育行业的应用相对较少.

本文以四川交通职业技术学院为研究对象, 分析了高职院校信息化现状, 讨论了在高职院校建设数据仓库的基本策略以及相关内容.

1 四川交通职业技术学院信息化现状

四川交通职业技术学院是全国第二批国家示范高职院校, 其信息化建设走在了全省高职院校的前列, 目前已建成了多个应用系统, 如: 教务管理系统、收费系统、学生工作管理系统、图书管理系统等应用系统. 这些应用系统在日常的业务中发挥了巨大的作用, 提高了部门的工作效率和学院的管理水平. 但是, 由于这些系统是在不同时期购买或开发的, 系统与系统之间没有统一的接口和标准, 系统与系统之间没有共享数据, 信息孤岛现象严重, 多个应用系统之间重复采集数据现象严重, 产生了严重的数据不一致.

现行的应用系统无法满足决策分析的需要. 决策分析通常所需的是面向主题的高层次的汇总数据, 其数据往往来源于多个应用系统, 由于信息孤岛的存在, 导致了决策分析难以实现; 历史的数据（例如一年前的数据）往往没有在业务系统中备份, 导致宏观分析、决策分析、长期历史分析难度很大. 因此需要使用数

① 收稿时间:2012-09-18;收到修改稿时间:2012-10-22

据仓库技术, 将各个业务系统的数据汇集在一起.

2 建设数据仓库的基本策略

根据数据仓库中所存储的数据和其所面向的主题, 从结构上来讲, 数据仓库主要有三种模型: 企业级数据仓库、数据集市和虚拟数据仓库[2]. 企业级数据仓库是指从整个企业的角度来考虑数据仓库的实施, 采用的是 “自顶向下”[3]的方法; 数据集市是指面向某一个部门或特定的主题开始构建数据仓库, 是一种特殊形式的数据仓库, 随着主题的增加, 逐渐涵盖整个企业, 采用的是 “自底向上”的方法; 虚拟数据仓库是利用中间件技术把分散在不同应用系统中最重要的数据集中到数据中心, 应用系统通过数据中心提供的 Web service 访问其他业务系统.

企业级的数据仓库, 投资大、建设周期长, 回报也慢; 虚拟数据仓库不是真正意义上的数据仓库, 其能有效解决信息孤岛问题, 但还是不能满足决策分析的需要; 数据集市采用 “自底向上”的方法, 投资相对较少, 建设周期较短, 其缺点是, 当增加新的主题时, 可能带来较多的数据冗余. 结合四川交通职业学院的信息化现状, 要把全部的数据统一起来有一定的难度, 而且用户对数据仓库的需求也不是非常明确, 因此采用数据集市是一种切实可行的方法.

3 确定主题

目前学院决策层最希望了解学生成绩和学费两方面的信息, 因此以教务管理系统和学生收费系统数据位基础, 定义成绩和学费两个主题, 以后再逐步扩展. 围绕每一个主题定义事实表和维度表. 主题主要信息描述如表 1 所示. 成绩和学费主题可以共用学生信息.

表 1 主题信息表述表

主题	主要属性
成绩	学生信息：学号、姓名、班级、专业、系部 课程信息：课程编号、课程名称、授课教师、开课时间等 成绩信息：分数等
学费	学生信息：学号、姓名、班级、专业、系部 缴费信息：实缴学费、应缴学费、缴费时间等

4 确定粒度

粒度是指数据仓库的数据单位中保存数据的细化或综合程度的级别, 数据越详细, 粒度就越小, 级别也就越低; 数据综合度越高, 粒度就越大, 级别也就越高[4]. 粒度会影响数据仓库逻辑结构的设计、数据存储和最终的分析效果. 从四川交通职业技术学院的实际情况来看, 数据量不是特别大, 因此粒度设计应遵循“最小粒度原则”[4]. 成绩信息和学费信息在教务管理系统和收费系统中记录到每一学期, 而目前针对成绩和学费的数据综合统计有分年度聚合和分学期聚合, 因此时间的最小粒度应为学期.

5 数据仓库逻辑模型设计

5.1 星型架构

星型架构中有两种类型的表: 事实表和维度表. 事实表用来存放度量, 维度表用来存放维度. 事实表和多个维度表通过主外键关系连接而成, 维度表与维度表之间没有关系. 事实表是整个星型架构的核心, 它包含了事实数据和维度表的关键字.

5.2 雪花型架构

雪花型架构使用层次结构的方式存储维度, 每一层都存储在单独的维度表中, 这些维度表的连接方式类似于雪花的形式, 因此叫雪花型架构. 通常情况下, 雪花型架构具有关系数据库设计的所有优点, 雪花型架构会将维度表规范化到第三范式[5], 减少了数据的冗余, 提高了数据的存储效率. 但是雪花型架构的不足之处在于, 当在层次结构的较上层进行聚合时, 需要较多的表连接, 通常数据仓库中的数据都比较多, 这可能导致性能的下降. 同时, 如果用户想要看到底层之上的维度时, 都需要进行临时聚合计算, 如果数据仓库中的维度较多或者维度有很多成员时, 也会耗费较多的时间.

5.3 逻辑模型设计

数据仓库包含两方面的主题, 因此需要建立两个事实表, 成绩事实表和学费事实表. 成绩事实表包含如下数据: 学号、学生所在班级编号、课程编号、时间编号、教师编号和成绩, 其中, 成绩作为度量值. 学费事实表包含如下数据: 学号、学生所在班级编号、时间编号、应缴学费和实缴学费, 其中应缴学费和实缴学费是度量值.

数据仓库还应该包含以下维度表: 时间维度表、学生维度表、部门维度、教师维度和课程维度. 时间维度包含学年和学期; 学生维度包含学号、姓名、性别、所在班级编号和入学时间; 班级维度包含班级编

号、班级名称和班级所在专业编号；专业维度包含专业编号、专业名称以及专业所在系别编号；部门维度包含部门编号和部门名称；课程维度包含课程代码和课程名称；教师维度包含教师编号、教师姓名、教师性别. 其中班级维度、专业维度和系别维度比较特殊，在做分析的时候，经常需要从系的层面下钻到专业层面，再从专业层面下钻到班级甚至下钻到学生，或者再反过来依次上钻，因此数据仓库采用了雪花型架构. 数据仓库逻辑模型如图 1 所示.

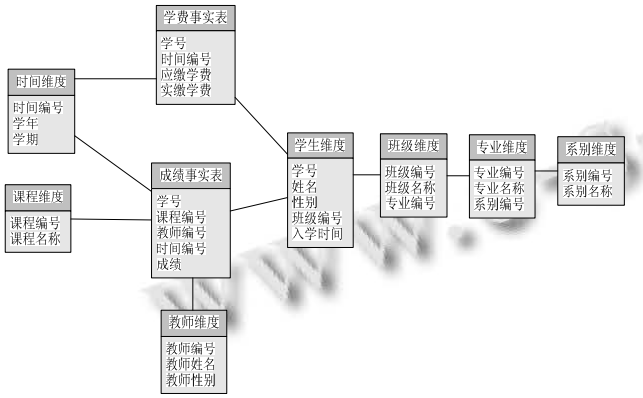

图 1 数据仓库逻辑模型

6 加载数据到数据仓库

数据仓库建立好以后，需要使用 ETL 工具将教务系统和收费系统中的数据加载到数据仓库. 但是这些数据可能有错误，为了提高进入数据仓库中数据的质量，需要将数据进行预处理才能加载到数据仓库中，因此最为关键的就是对数据进行清洗.

在教务管理系统中，数据质量问题主要是有部分课程的授课内容相同，在不同年级开设，分别有不同的名称. 如计算机系大一学生开设的一门课程《程序设计基础》存在《程序设计基础 4》、《程序设计基础 6》以及《程序设计基础与编程逻辑》等多个课程名称，需要将课程名称都统一成为《程序设计基础》. 首先使用条件性拆分组件对数据进行拆分，将课程名称为《程序设计基础》的数据进行拆分，流程如图 2 所示.

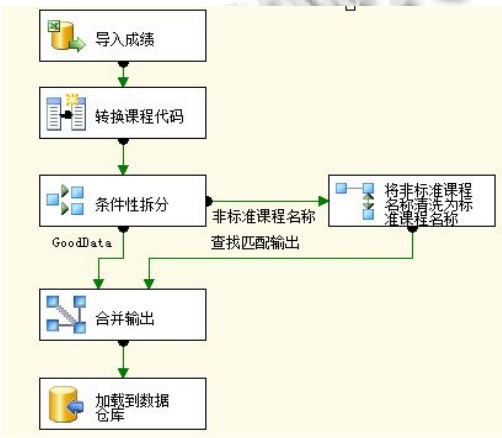

图 2 清洗课程名称示例

7 基于数据仓库的OLAP分析

7.1 学生学年学期平均成绩趋势分析

对于一个学生的评价与分析，不能只是依靠一门或两门课程的成绩，应当从学生的成绩稳定性、是否有进步或者退步、成绩波动的幅度、变化趋势等多方面综合进行评价与分析，才是比较科学的分析方式. 笔者选择了三个同学的成绩，观测其从入学到毕业成绩变化趋势以及成绩波动情况，如图 3 所示. 从结果可以看出，三个同学的平均成绩在大一的下学期都出现了下滑，然后再呈现上升趋势，因此我们可以去寻找造成同学们成绩下滑的原因.

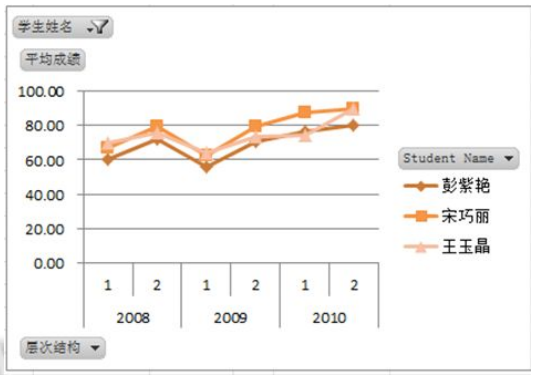

图 3 学生成绩趋势分析

7.2 同一门课程在不同年级平均成绩趋势分析

对于年级相差不远，同一专业的学生来讲，大学期间所学习的课程基本是相同的. 通过分析同一门课程在不同年级的平均成绩，可以了解学生的整体素质，分析结果如图 4 所示. 从分析结果可以知道，BCIT08 级和 BCIT10 级程序设计课程《C 语言的编程程序在交通中的应用》平均成绩都比 BCIT07 级和 BCIT09 较低，因此，我们知道对于这两个班级的同学在以后上软件编程方面的课程的时候，尤其要注意授课内容和教学方法等，加强理论知识，注重学生技能训练，提高学生编程的能力.

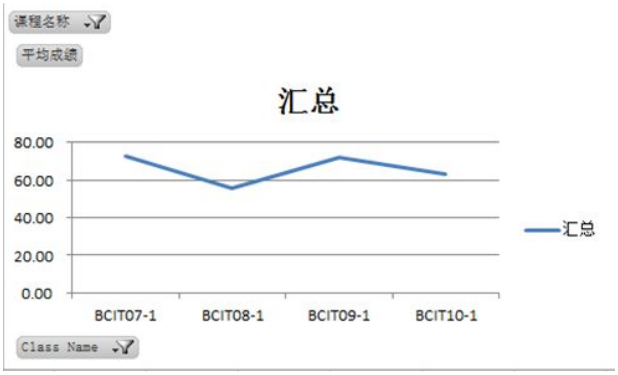

图 4 同一门课程在不同年级平均成绩趋势分析

7.3 欠费构成分析

对于决策者来讲，往往想知道欠费的构成，例如哪些系、班级欠费了，哪些系、班级欠费比较多，可以采用 OLAP 的上卷、下转操作实现，分析结果如图 5 所示.

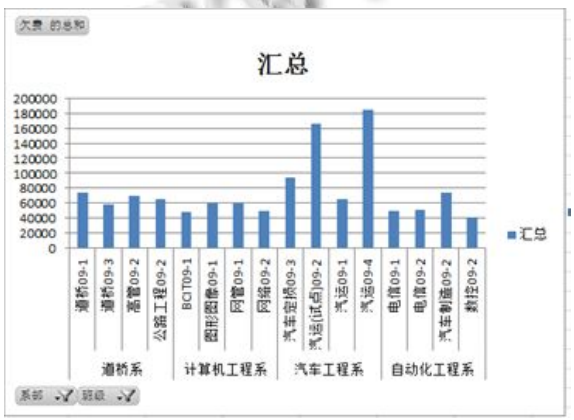

图 5 同一门课程在不同年级平均成绩趋势分析

8 结语

作为一名高职院校教育工作者，笔者深刻的体会到，高职院校信息化建设相对于本科院校来讲相对滞后，多年以来在教务管理系统、学生收费系统中积累了大量的数据，但是由于教务管理系统、学生收费系统等应用系统偏重于数据的采集与管理，难以对数据进行深入发掘，提供客观的、有用的信息. 笔者所在的单位数据仓库建立于两年前，通过这两年的运行，发现采用数据仓库技术有效的提高了教务系统和学生收费系统信息的集成度，有效的避免了信息孤岛；基于成绩和学费两个主题的 OLAP 分析有效的为学院决策者、系部管理者以及老师提供了可靠的数据支持，从多个角度了解学生的状况；同时也发现，在其他方面如学院的招生情况、学生借阅图书的情况等也需要进行深入分析，这就需要在数据仓库中完善相关主题，形成相对完整的数据仓库.

参考文献

1 Inmon WH. Building the Data Warehouse. New York: John Wiley and Sons Inc.1990,48–50.
2 袁连海.数据仓库技术研究与实现.成都:西南交通大学,2002.
3 马国俊.基于 OLAP 的企业数据仓库规划与建设.制造业自动化,2011(12):56–59.
4 朱德利. SQL Server 2005 数据挖掘与商业智能完全解决方案.北京:电子工业出版社,2007.100–101.
5 赵志恒,武海锋.Microsoft SQL Server 2005 商业智能实现.北京:清华大学出版社,2008.200–202.

(上接第 131 页)

5 Jayaram S, Schmugge S, Shin MC, Tsap LV. Effect of colorspace transformation, the illuminance component, and color modeling on skin detection. American: IEEE, 2004: 813–818.
6 Kwok NM, Fang G, Ha QP. Effect of color space on color image segmentation. American: IEEE, 2009:1–5.
7 成春喜,全燕鸣.基于 HIS 模型的彩色图像背景减法.计算机应用,2009,29:231–232,235.
8 黄志勇,孙光民,李芳.基于 RGB 视觉模型的交通标志分割.微电子学与计算机,2004,21(10):147–148,152.
9 Wang S. Color image segmentation based on color similarity. American: IEEE, 2009:1–4.
10 Chapron M. A new chromatic edge detector used for color image segmentation. Washington: IEEE,1992: 311–314.