文章编号：2095-2163(2019)03-0122-04 中图分类号：TP315 文献标志码：A

数据仓库技术在高校数据统计与分析系统中的应用研究

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：为更好地支持各业务系统数据的整合集成以及对多维数据的交叉分析和数据的挖掘分析等需求，有效支撑高校管理与决策对数据的需求，需搭建规范、有效的数据仓库平台，并部署实施数据统计与分析系统。基于此，文章阐述了数据仓库的基本理论及一般构建方法，给出了符合高校需求的数据仓库主题划分，设计了 5 层数据仓库结构体系。并将其应用于湖南铁道职业技术学院数据分析系统，取得了较好的效果，对其它高校数据仓库的构建具有示范和借鉴意义。

关键词：数据仓库；数据分析；ETL

Research on the application of data warehouse technology in university data statistics and analysis system

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

【Abstract】 In order to better support the integration of data of various business systems and the cross-analysis of multi-dimensional data and data mining and analysis, and effectively support the data management needs of university management and decision-making, it is necessary to build a standardized and effective data warehouse platform and deploy and implement a data statistics and analysis system. Based on this, the article expounds the basic theory and general construction method of data warehouse, and gives the classification of data warehouse subject to the needs of colleges and universities, and designs a five-tier data warehouse structure system. It is applied to the data analysis system of Hunan Railway Professional Technology College, and has achieved good results. It has demonstration and reference significance for the construction of other university data warehouses.

【Key words】 data warehouse; data analysis; ETL

0 引 言

随着信息技术的迅猛发展，教育信息化的建设在大多数的高校里已经开展多年，用于支持高校管理及教育教学的多个信息系统已经完成建设，大量的业务数据已经产生，如何有效地组织管理这些数据，使其能为学校的管理决策和教育教学提供支持已成为当前高校教育信息化研究的重要课题。教育部于 2018 年 4 月印发的《教育信息化 2.0 行动计划》中就明确指出：深化教育大数据应用，全面提升教育管理信息化支撑教育业务管理、政务服务、教学管理等工作的能力。充分利用云计算、大数据、人工智能等新技术，构建全方位、全过程、全天候的支撑体系，助力教育教学、管理和服务的改革发展[1]。

数据仓库技术能够从数据采集、清洗、存储、统计分析、展现等多个方面为高校数据资源的积累、管理和利用提供技术方案。数据仓库中的数据为分析型数据，以相同主题的方式组织聚集的，而业务系统数据库中的数据为操作型数据，是围绕着一个或几个业务处理流程来组织。所以，利用数据仓库技术来建设高校数据统计与分析系统是必要且有效的。本文根据高校数据统计与分析的实际业务需求及教育数据的特点，结合现有数据仓库的相关技术和建设经验，设计符合高校数据统计与分析的数据仓库架构，给出建设过程，最后实现基于数据仓库的高校数据统计与分析系统。其系统架构及建设过程可为其它类似数据分析或决策支持系统的建设提供有益参考和借鉴。

1 数据仓库基本理论

数据仓库之父 Bill Inmon 给出的定义是：数据仓库是一个面向主题的（Subject Oriented）、集成的（Integrate）、相对稳定的（Non-Volatile）、反映历史变化（Time Variant）的数据集合，用于支持管理决策[2]。其中的主题是指用户使用数据仓库进行数据分析时所关心的重点方面。每一个主题对应一个

基金项目：2018 年度湖南铁道职业技术学院校级课题（K201823）；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：大数据、数据库技术。

收稿日期：2019-03-10

宏观的分析领域。数据仓库排除对于决策无用的数据，提供特定主题的简明视图。集成就是对原有分散的数据进行抽取、清洗，然后进行加工、汇总、整理得到的规范、统一的全局信息。同时数据仓库中的数据是很少有修改或是删除操作的，数据将会长期保留，包含有大量的历史数据，能够存储不同时间范围的数据快照，所以这就既是相对稳定的，也是能够反映历史变化的数据集合。

因为数据仓库的目的就是数据统计分析，为决策支持服务，这就注定数据仓库不会“生产”数据，其内部数据主要来源于其它业务系统或是外部数据，这些数据经过清洗、转换流入至仓库层中。在仓库层中，有用于对仓库中数据进行描述管理的元数据；有为应用层提供在线分析处理服务的 Web 服务器；有面向各类的主题的聚集数据；有面向特定应用或单个主题的数据集市等。最后，由具体的应用为用户提供各类数据统计分析等相关服务。由此，研究可以将数据仓库分为 3 层：源数据、仓库层、应用层，如图 1 所示。

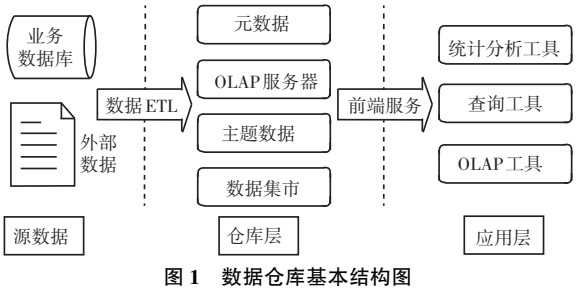

图 1　数据仓库基本结构图

Fig. 1　Data warehouse basic structure

2　系统构建

2.1　源数据分析及主题确定

建立数据仓库的首要环节就是对业务源数据进行分析，是否能够充分理解和透彻分析业务源数据将直接影响到后面数据模型的设计是否合理、有效。通常，研究将源数据的分析分为表级分析和字段级分析。而源数据分析的主要工作则包括有：整理所有业务系统数据库的清单表，统计每一张数据表的记录数，对业务系统数据库的数据量以及使用情况有一个全面的了解；统计每一张表的主要字段，包括数据类型以及数据长度；了解表的主键和外键，明确表之间的相互依赖关系；明确最后需要进入数据仓库的数据表及具体数据项。

主题是在较高层次上将业务系统中的数据进行整合、归类和分析利用的一个抽象概念，每个宏观的分析领域与一个主题相对应。在逻辑意义上，主题是对应业务中具体宏观分析领域所涉及的分析对象。主题一般是采用逐级细分的思路进行设计划分，结合高校的具体业务数据特点以及数据统计与分析系统的建设需求，本次研究可以开发确定的数据仓库主题域即如图 2 所示。

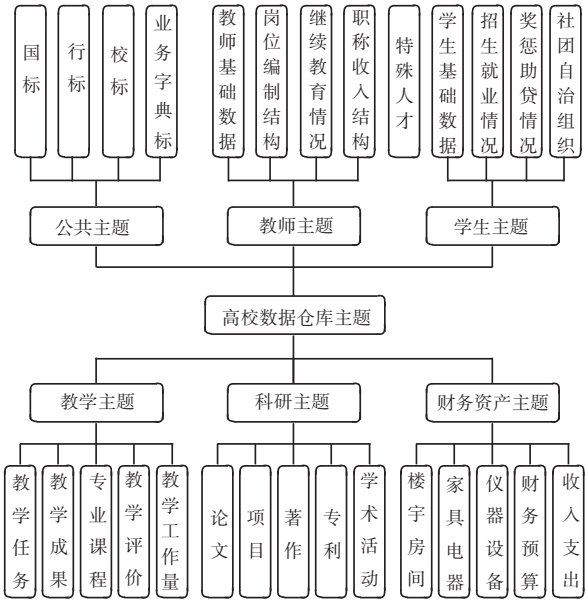

图 2　数据仓库主题域

Fig. 2　Data warehouse subject domain

2.2　分层架构设计

数据统计与分析系统的核心部分就是数据仓库，数据仓库的关键任务就是对分散、繁杂的数据进行规范、统一的清洗处理，然后建立清晰的业务逻辑关系，实现数据的汇总聚合，分层的架构设计能够有效地完成上述工作。结合高校对数据仓库的实际业务需求，将数据仓库按照相关功能分为临时层、近源层、主题层、汇总层和集市层等 5 层，其层次架构设计如图 3 所示。对其中各层的功能设计可做阐释解析如下。

(1)临时层：该层位于最下层，最接近业务系统数据库，并直接与各个业务系统数据源对接，尽量保持业务数据的原貌，在抽取策略上可选择增量和全量抽取，同时在抽取时加上时间戳，形成多个版本的历史数据信息。

(2)近源层：该层以偏源系统建模，对临时层数据进行初步的筛选和加工，不进行数据的整合处理，能够提供基于业务数据的访问需求。

(3)主题层：该层数据模型要求符合数据库 3NF 范式规则，是数据仓库的核心数据层，每个主题对应一个宏观分析领域。主题层数据强调数据整合和历史信息，能够支持较长时间周期的分析需求，模型设计应具备足够的灵活性，以满足进一步的升级和更新。

(4)汇总层:该层是面向数据分析而建模,由于主题层是大量高度规范化的数据,因此要完成一个查询或是一项统计往往需要大量的关联工作[3],而数据统计分析又需要针对主题层数据进行大量的汇总工作(如:累加、平均值、记录数、最大最小值等等),设计汇总层能有效提高数据仓库的查询效率。

(5)集市层:该层数据面向具体应用,可以理解为供用户直接访问。具有面向用户、相互独立、形式各异等特点。同时,该层数据的生命期是依具体应用的需求而定。

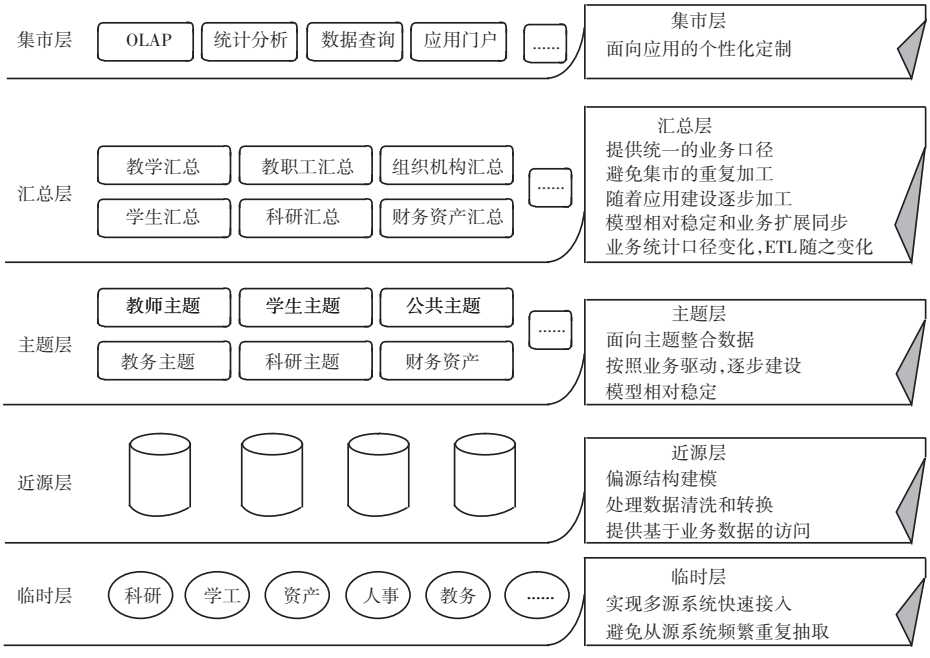

图 3 层次架构设计

Fig. 3 Hierarchical architecture design

2.3 数据 ETL 设计与实施

ETL(Extract-Transform-Load)是数据源端不同数据库或异构数据源的数据经过抽取、转换和加载到目的端的过程[4],ETL 是实施数据仓库的核心和灵魂。ETL 根据不同的抽取策略进行参数、执行时间的初始化,之后进入后台守护模块实时监控。后台主要完成将业务数据按照抽取策略定时导入数据到临时数据库,处理后定时调用后台存储过程保存到数据仓库中。ETL 数据处理整体上可分为数个阶段。对此可做设计分述如下。

(1)数据抽取阶段。主要根据业务调研制定的规范,把分布在各业务系统的数据抽取到数据仓库的临时层,在抽取时要遵守业务调研阶段制定的数据标准。

(2)数据清洗阶段。是解决从各数据源抽取数据所出现的数据重复、数据不一致、空值数据等问题,包括标准化处理、空值处理和不一致数据处理等。数据清洗工作主要在近源层完成,清洗时结合一定的业务规则将数据值进行标准化。

(3)数据转换阶段。也是在近源层完成,即将源系统抽取的数据,经过不同的算法处理,并将数据处理成数据仓库特定的存储模式。数据转换的任务主要是进行不一致的数据转换、数据粒度转换和一些业务规则的计算。

(4)数据加载阶段。就是在数据仓库中经过数据转换、清洗后,按一定的方式把数据存储到数据仓库的主题层、汇总层和集市层中。

常见的 ETL 工具有 Informatica、Datastage、ODI、OWB、kettle 等,本文所选取的是 ORACLE 公司的 ODI 系统。ODI 提出了知识模块的概念,用户既可以直接使用 ODI 的知识模块完成数据的获取工作,也可以直接在知识模块上面做各种定制,比如某一个业务场景可能并不需要知识模块里的某一个特定的步骤,那就可以直接把该步骤删除掉从而提供更好的性能。当然用户也可以完全自己来开发这些知识模块[5]。

2.4 数据展示

通过对各类数据的分类汇总,形成各个主题的事实数据表,利用数据展示工具集中提供统计数据的图表展示、查询下载等相关服务。常见的用于数据前端展示的工具有 IBM 的 Cognos 和 SAP 的 BO、Oracle 的 BIEE、微软的 SSRS 等,这些工具都是专业的报表工具,属于商业 BI 产品,具有强大的报表设计能力和丰富的

数据展示功能，但是却也价值不菲、且使用较为复杂。ExtJS 是一款基于 JavaScript 的前端用户界面开发平台。作为一款前端 AJAX 框架，ExtJS 具有编程简单、与后台技术无关、功能强大等特点，可以便捷地将 ExtJS 用在 dot NET、Java、PHP 等各种设计语言开发的项目中。基于数据仓库的高校数据统计与分析系统就是基于 ExtJS 开发。图 4 展示科研数据统计分析报表的详细界面，这是系统的典型展示界面，上面是按相关维度生产的数据统计图和数据汇总表、当点击图表中某个具体统计数字时，则在页面下方显示划分至更细粒度的详细数据，同时提供查询和导出功能。

图 4　数据统计展示界面

Fig. 4　Data statistics display interface

3　系统实现

数据仓库平台的物理结构主要包括数据交换中心服务器、数据仓库服务器、数据管控中心服务器、数据备份服务器。平台系统的部署设计如图 5 所示。其中，数据交换中心主要负责各业务数据库与数据仓库之间的数据抽取、清洗和推送，部署 ODI 系统。数据管控中心负责对各类数据标准、数据接口的维护和管理；数据备份服务器主要负责对仓库数据的容灾备份。数据统计与分析系统服务器可同时担任 Web 应用服务器和数据库服务器。在此部署方案中，充分考虑了平台系统的整体稳定性和可用性。平台的各功能服务器尽量独立，保证了系统的运行效率和数据安全。

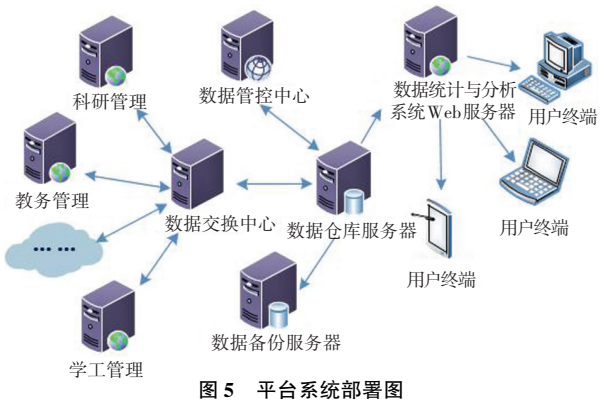

图 5　平台系统部署图

Fig. 5　System deployment diagram

4　结束语

数据仓库平台的构建是基于高校对数据分析业务的实际需求建立的，完全依赖于高校各应用系统的业务运行情况，各个高校的业务需求及定义都不完全一样，同时还有动态调整的潜在需求。所以，构建出来的数据仓库系统也都各具特色。本文以湖南铁道职业技术学院数据分析系统的构建为例，简要介绍了数据仓库的一般构建过程，研究剖析了一种数据仓库平台的架构设计机理，同时一并给出了该设计平台的部署方案。该系统达到了既定需求，能够灵活地对各类业务数据进行整合分析，形成各类数据报表，为学校的相关的管理决策提供数据支撑。

参考文献

[1] 中华人民共和国教育部. 教育部关于印发《教育信息化 2.0 行动计划》的通知[EB/OL]. [2018-04-18]. http://www.moe.edu.cn/srcsite/A16/s3342/201804/t20180425_334188.html.
[2] 吴振涛. 基于数据仓库技术的数据集成在数字化校园中的应用[J]. 电子设计工程，2016，24(9)：28-31.
[3] 龙新征，李丽，彭一明，等. 基于数据仓库的高校数据统计服务平台研究[J]. 通信学报，2013，34(Z2)：163-169.
[4] 张孟春. 面向数据集成的分布式 ETL 研究与设计[J]. 软件导刊，2017，16(11)：197-199.
[5] 李欣. 高校信息门户内外网信息流转系统的设计与实现[D]. 成都：电子科技大学，2012.

(上接第 121 页)

参考文献

[1] 米增强，王飞，杨光，等. 光伏电站辐照度 ANN 预测及其两维变尺度修正方法[J]. 太阳能学报，2013，34(2)：251-259.
[2] 付青，单英浩，朱昌亚. 基于 NARX 神经网络的光伏发电功率预测研究[J]. 电气传动，2016，46(4)：42-45.
[3] 丁明，王磊，毕锐. 基于改进 BP 神经网络的光伏发电系统输出功率短期预测模型[J]. 电力系统保护与控制，2013，40(11)：93-99，148.
[4] 倪春华，陈国恩，朱伟，等. 基于相似日理论和 BP 神经网络的光伏发电功率预测[J]. 电力应用，2016，35(1)：42-48.
[5] 徐潇源，严正，冯冬涵，等. 基于输入变量秩的相关系数的概率潮流计算方法[J]. 电力系统自动化，2014，38(12)：54-61.
[6] 刘俊，王旭，郝旭东，等. 基于多维气象数据和 PCA_BP 神经网络的光伏发电功率预测[J]. 电网与清洁能源，2017，33(1)：122-129.
[7] OUDJANA S H，HELLAL A，MAHAMED I H. Short term photovoltaic power generation forecasting using neural network [C]//2012 11th International Conference on Environment and Electrical Engineering. Venice，Italy：IEEE，2012：706-711.
[8] BIZZARRI F，BONGIORNO M，BRAMBILLA A，et al. Model of photovoltaic power plants for performance analysis and production forecast[J]. IEEE Transactions on Sustainable Energy，2013，4(2)：278-285.