责任编辑　沈晓平　楼玲玲
封面设计　赵　军

世纪出版

高校数据仓库系统建设与应用

University Data Warehouse

www.sstp.cn

上架建议：大数据·计算机技术

ISBN 978-7-5478-3462-6

定价：85.00元

易文网：www.ewen.co

高校数据仓库系统建设与应用

陈　云　主编

上海科学技术出版社

内容提要

全书共分 6 章。第 1 章主要介绍数据仓库的概念、背景及高校大数据研究展望。第 2 章介绍数据仓库相关的技术原理。第 3 章介绍数据仓库项目需求分析的流程、方法、原则，以及数据仓库需求分析应包含的内容。第 4 章结合高校业务特点，根据数据仓库系统建设过程介绍其各个环节，主要有建设流程、系统架构、主题规划、源数据分析、数据模型设计、ETL 设计、ETL 开发、ETL 测试、系统运行监控和报表设计开发，重点阐述每个环节采用的技术方法、实现过程及注意事项。第 5 章介绍数据仓库元数据管理、数据质量管理和系统运维管理，其中重点介绍数据质量监控系统的设计及实现。第 6 章通过剖析高校的核心管理业务、数据分析指标和分析案例，详细说明数据仓库在高校各业务管理中的应用。

智能决策支持应用是高校信息化建设发展到高级阶段的必然产物，本书可供高校信息化建设主体单位、高校信息化主管部门、高校决策层、管理人员、IT 人员参考。

前　言

大数据时代的来临，数据已经渗透到当今每一个行业和业务职能领域，成为重要的生产因素。随着高校信息化的发展，学校教学、科研、后勤等管理和服务系统建设初具规模，经过多年建设和应用，规范了业务管理流程，提升了业务管理效率，也积累了大量数据资源。随着高校从规模扩张转向内涵式发展，如何合理配置资源、实现教育资源共享，使资源能最大限度地发挥效用，已成为高校在激烈竞争中需要解决的重要问题。高校面向管理与服务的信息化已经不再满足于事务处理和信息查询，而需要利用积累的大量数据进行相关分析，将其转换成有用的信息和知识，为学校管理和决策提供有力支撑，提升高校数据服务的能力。所以如何利用这些宝贵的数据资源为高校管理与决策提供服务，已成为目前高校信息化建设的重要课题。

数据分析服务、智能决策支持应用是高校信息化发展的高级应用阶段，众多高校在这方面已经开始了探索，并且取得了初步成效，但是由于建设方法和过程不规范、系统架构不合理、应用需求不明确、核心技术瓶颈等问题，导致高校的数据服务和智能决策支持应用普遍较弱。而数据仓库是智能决策支持应用的基础平台，上海财经大学信息化办公室从 2008 年开始，在高校数据仓库建设方面进行了深入探索和实践，积累了较为丰富的经验，取得了较为显著的应用成效。本书定位是高校信息化建设单位实施数据仓库项目的指导类书籍，重点介绍高校数据仓库建设与实施过程中的相关内容，包括数据仓库项目的需求分析该如何开展、高校数据仓库设计与开发、数据质量管理，以及核心业务应用案例等内容。

本书针对目前高校建设数据仓库的普遍问题，结合上海财经大学数据仓

库建设的实践，通过深入分析高校数据服务需求，提出基于数据仓库的高校数据服务系统技术方案，探索适合高校业务特点的数据服务系统实施方法，形成高校数据服务系统实施和应用的典型案例，从而推动和促进数据服务在高校中的广泛应用。全书主要从数据仓库需求分析、设计、开发、管理与应用几个方面进行阐述。

本书由上海财经大学信息化办公室组织撰写，多位作者参与本书撰写。其中，第1章由王珊珊、陈云、吴先斌撰写，第2章由吴先斌撰写，第3章由王珊珊撰写，第4章由高亮撰写，第5章由吴先斌、高亮撰写，第6章由高亮、吴先斌、陈云撰写。全书由陈云总撰定稿。同时，感谢项目合作单位上海吉贝克信息技术有限公司提供的相关资料。

由于作者水平有限，书中错误在所难免，恳请专家和广大读者批评指正。

编　者

目　录

第1章 绪 论

1.1 数据仓库概念

数据仓库的定义被广泛接受的是 William H. Inmon 在《Building the Warehouse》一书中给出的，他认为数据仓库是一个面向主题的(Subject Oriented)、集成的(Integrated)、非易失的(Nonvolatile)且随时间变化(Time Variant)的数据集合，用来支持管理人员的决策分析。总之，对于数据仓库的概念，可以从两个层次理解：首先，建设数据仓库系统的目的在于支持分析决策，面向分析型数据处理，它不同于企业现有的操作型数据库；其次，数据仓库是对多个异构数据源的有效集成，集成后按照主题进行存储，并包含历史数据，而且存放在数据仓库中的数据一般不再修改。根据上述数据仓库的概念可以总结出数据仓库具有以下四个特点。

1）面向主题性

面向主题是数据仓库中组织数据的基本原则，数据仓库中所有数据都是围绕某一主题组织、展开的。主题是一个抽象概念，是在较高层次上将企业信息系统中的数据综合归类并进行分析利用的抽象。在逻辑意义上，它对应企业中某一宏观分析领域所涉及的分析对象。从信息管理角度看，主题就是在一个较高的管理层次上对信息系统中的数据按照某一具体的管理对象进行综合、归类所形成的分析对象。数据仓库中面向主题组织数据与传统数据库面向应用组织数据的特点相对应。例如，高校数据仓库所组织的主题有教

学分析、科研分析、资产分析等，而按照应用系统来组织则会是教学系统、财务系统、科研系统、人事系统、资产系统等。主题组织是按照分析的要求来确定的，这与按照应用的要求组织数据不同。例如，同样是教师科研，在操作性数据库系统中用户关心的是如何方便地进行科研项目申报、项目验收、科研成果管理；而在进行分析处理时，用户关心的是科研的投入产出情况、各类科研项目完成情况、科研水平等。

2）集成性

数据仓库的集成性是指根据决策分析的要求，将分散于各处的源数据进行抽取、筛选、清理、综合等集成工作，使所有数据能够统一、有机地整合在一起。数据仓库每个主题所对应的源数据通常来源于不同的数据源（如关系数据库、一般文件和联机事务处理记录），这些数据只为业务日常处理服务而不是为决策分析服务，而且这些数据往往具有重复性和不一致性。所以，首先要从源数据中挑选出数据仓库所需要的数据，将这些数据按照标准进行统一，确保命名约定、编码结构、属性度量的一致性，然后再将原始数据结构从面向应用到面向主题进行转变。以高校为例，支撑高校运营管理的信息系统一般包括：教学管理系统、人事管理系统、科研管理系统、财务管理系统、资产管理系统等，如果要进行高校教师科研分析、教学情况分析、学生学习情况分析、资产使用情况分析等，就可能需要将业务系统的数据进行集成、整合才能完成。高校信息化建设存在的困难和问题是由于信息系统较多，信息系统由不同的开发团队、在不同时间进行开发实现的，导致数据标准和技术平台很难统一，如学科、专业等数据有教育部和科技部两种标准，而实际情况是并非所有的信息系统都参照标准，在数据仓库建设过程中需要花费很大的精力将这些数据按照标准进行统一。因此，在高校信息化建设过程中统一数据标准是件非常重要的事情。

3）非易失性

在数据仓库中，数据是从联机事务处理系统中抽取出来，存储一段相当长时间内的历史数据，是不同时间点数据快照的集合，以及基于快照的统计、综合和重组。数据仓库中的数据主要为企业提供决策分析服务，所涉及的数据操作主要是数据查询，一旦数据进入数据仓库，只要数据没有超过数据仓库的数据存储期限，一般不对数据进行更新和删除操作，只进行查询。数据

仓库的数据的存储期限主要由数据分析对数据量的需求，以及服务器的运算能力、存储能力等因素综合决定。

4) 时变性

数据仓库的时变性就是数据应该随时间的推移而发生变化，尽管数据仓库中的数据不像业务数据库那样要反映业务处理的实时状况，但是数据也不能长期不变，如果依据10年前的数据进行分析，那决策所带来的后果将是十分可怕的。因此，数据仓库能够不断捕捉主题变化的数据，将那些变化的数据加到数据仓库中，也就是说在数据仓库中不断生成主题的新快照，以满足决策分析的需要。新数据快照的间隔根据快照的生成速度和决策分析的需要而定，可以是一小时、一天，还可以是一周。快照是业务处理系统的某一时间的瞬态图，而这些瞬态图则构成了数据仓库中数据的不同画面，这些画面的连续播放可以产生数据仓库的连续动态变化图，这十分利于高层管理的决策。数据的时变性不仅表现在数据的追加上，还反映在数据的删除上。尽管数据仓库中的数据不像业务系统中的数据那样只能保留数月，但是在数据仓库中数据的存储期限还是有限的，一般考虑保留5～10年。

1.2 高校数据仓库建设背景与意义

随着高校信息化的迅速发展，学校教学、科研、后勤等管理和服务系统建设初具规模，经过多年建设，规范了业务管理工作，提升了业务管理水平，也积累了大量业务数据。随着高校从规模扩张转向内涵式发展，如何合理配置资源、实现教育资源共享，使资源能最大限度地发挥作用，已成为高校在激烈的竞争中需要解决的首要问题。因此，高校的管理与服务的信息化已经不再满足于对数据单纯的处理和查询，而需要利用这些收集到的大量数据进行相关分析，将其转换成有用的信息和知识，为决策提供有力的支撑，进而提升高校数据服务的能力。如何利用这些宝贵的数据资源为高校管理与决策提供服务，已成为目前高校信息化建设面临的重要难题。

(1) 目前高校数据服务存在的主要问题有以下几方面：

① 数据服务应用实施不成体系。各高校开始注重应用多维分析以及数据挖掘技术对数据进行分析，但研究往往过于片面、不够深入，没有形成完整

的、可行的理论体系和解决方案。

② 实施高校数据服务需要克服一系列的技术困难。要在现行的管理信息系统基础上加以改造，实现数据分析和辅助决策是相当复杂的。首先是由于现有系统中存在着许多针对不同应用所定制的专用系统，它们运行在不同的平台之上，具有源头多样化、数据堆积无序、主体不突出、数据结构不同、数据采集困难、数据处理复杂等一系列问题；其次由于历史的数据量很大，不同系统的数据难以集成，对大量数据的访问及处理性能明显下降，系统运行的开销很大；最后往往应用系统是以事务管理为主，要在其上开发分析与辅助决策功能比较困难。

③ 基于数据服务的应用需求不够明确，成熟的应用比较少。虽然各高校已逐步开展各类数据分析，但往往由于需求分析不充分、方法不到位，使得数据应用无法满足用户需求。

如何整合不同系统的数据，形成统一的模式以方便处理？如何处理这些信息，让这些长期积累下来的财富不被白白浪费？如何利用现代化信息管理工具，从中挖掘出隐含的规律和方法，为决策者提供有力的支持？这些都是需要研究解决的关键问题。

数据仓库是一种能把各种源数据库集成一个统一的目标数据库，并能把各种数据转换成面向主题的格式，能从异构的数据源中定期抽取、转换和集成所需要的数据，便于最终用户访问，并能从时间维度进行分析，最后做出分析与决策的信息管理技术。

随着高校信息系统建设的日趋成熟以及管理要求的不断提升，可以引入数据仓库技术对高校信息系统数据进行结构重组，针对高校的特点和发展需求，按更有利于分析决策的角度去设计，构建以数据仓库为核心的数据中心，优化数据的存储和使用，在数据仓库之上进行数据挖掘等分析，让这些宝贵的数据资源实现真正的信息价值，提高对管理信息的利用率，为高校持续发展提供科学的决策依据，为学校的重大决策提供信息参考，进而提升高校整体管理和决策水平。

(2) 基于数据仓库的特点与优势，高校搭建数据仓库的作用体现如下：

① 支撑多维分析。数据仓库储存了大量的历史数据，可以通过分析不同时期和趋势对未来进行预测，这些数据通常不能被存储在一个事务型的数

据库里。学校管理人员可以不再仅凭有限的数据或他们的直觉，而是充分利用应用系统的各种数据做出分析与决策。

② 支持深度挖掘。在数据仓库基础上，可以进一步采用数据挖掘技术，对数据进行深度剖析。数据挖掘方法通常可以分为两大类：一类是统计型，实现相关分析、聚类分析和判别分析等；另一类是机器学习型，通过训练学习大量的样本集得出需要的模式或参数。

③ 提高数据质量。数据仓库的实施过程包括将数据从众多的业务系统中清洗并转换成共同的格式，大大提升了数据的准确性和有效性，而高质量的数据才是分析决策的重要基础。

④ 提高访问效率。用户可以快速访问许多数据源，而不用浪费时间从多种数据源中检索数据，同时海量数据的查询速度也得以保证。不仅如此，数据的查询也不会对业务系统运行造成影响。

上海财经大学已完成了覆盖各业务数据的数据仓库搭建，并在此基础上充分利用数据分析和数据挖掘技术构建了校务决策支持系统，既能为各类人员提供具体数据内容，又能从宏观角度反映学校发展变化，通过分析与挖掘高校历史数据，发现其中潜在的、深层次的、有价值的信息，以及内在关系和问题，抓住并解决关键问题，推动学校各项事业发展。

1.3 高校进入大数据时代

近年来，大数据在各行各业中掀起了巨大的风波，人们都在了解大数据，并思考如何利用大数据。随着云技术、移动技术、物联网技术在高校的应用，云学习平台、云图书馆、移动学习与服务、节能平台、智能监控等逐渐引入到学校的应用中，因此数据成倍增长，高校亦正进入“大数据”时代。

1.3.1 大数据概念与特征

目前，大数据还没有一个公认的定义，IDC（国际数据公司）认为大数据应当具有价值性，大数据的价值往往呈现稀疏性的特点。维基百科将大数据定义为：利用常用软件工具捕获、管理和处理数据所耗时间超过可容忍时间的数据集。在维克托·迈尔编写的《大数据时代》中，大数据是指不用随机分析

法，而采用所有数据进行分析处理的数据。研究机构 Gartner 将大数据定义为：需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产。综上，尽管定义不同，但基本都是从大数据的特征出发给出的。

大数据有以下四个层面的特征：

(1) 数据量较大(Volume)，一般大数据都拥有 PB 级别的量，当前典型个人计算机硬盘的容量为 TB 量级，而一些大企业的数据量已经接近 EB 量级。

(2) 数据类型多(Variety)，相对于以往便于存储的以文本为主的结构化数据，非结构化数据越来越多，包括网络日志、音频、视频、图片、地理位置信息等，这些多类型的数据对处理能力提出了更高要求。

(3) 处理速度快(Velocity)，这是大数据区分于传统数据处理的最显著特征，预计到 2020 年全球数据使用量将达到 35.2 ZB，在如此海量的数据面前，处理数据的效率就是企业的生命。

(4) 价值密度低(Value)，价值密度的高低与数据总量大小成反比，如何通过强大的算法更迅速地完成数据的价值"提纯"，成为目前大数据背景下亟待解决的难题，只有在合理利用数据并对其进行正确、准确的分析时，才能带来高价值的回报。

大数据最核心的价值在于对于海量数据进行存储和分析。相比现有的其他数据分析技术而言，大数据的"廉价、迅速、优化"这三方面的综合成本是最优的。

1.3.2 大数据关键技术

大数据技术是从各种各样不同类型的数据中快速获得有价值信息的能力。大数据来源于互联网、管理系统和物联网等信息系统，经过大数据处理系统的分析挖掘产生新的知识，用于支撑决策或业务的自动智能化运转。从数据在信息系统中的生命周期来看，大数据从数据源到最终应用需要经过数据准备、数据管理、计算处理、数据分析、知识展现五个环节，每个环节都不同程度面临技术上的挑战。大数据关键技术有：

1) 文件系统

该系统提供最底层存储能力的支持，是支撑上层应用的基础。Google

公司最早开发出了自己的文件系统 GFS(Google File System),随后微软也开发了自己的 Cosmos,随后又出现了 HDFS。

2) 数据库系统

由于关系模型的分布式数据库不能应对大数据时代大规模的压力,相应地提出了许多新型数据库系统,如 Google 的 Bigtable,Amazon 的 Dynamo 等,直到现在形成统一的 NoSQL。虽然 NoSQL 没有标准的定义,但一般认为具有模式自由、简易备份、最终一致性、支持海量数据等特征,同时形成了对应的索引与查询技术。

3) 数据分析

目前,最著名的计算模型为 Google 的 MapReduce。Google 公司目前针对 MapReduce 离线处理模式的不足,提出了基于 Web 数据级别的交互式数据分析系统 Dremel,能够实现极短时间内的海量数据分析。在离线与实时处理模式上,已经出现了二者融合的趋势。

4) 大数据处理工具

Hadoop 是目前最为流行的大数据平台,目前对该平台进行改进以便应用到各种场景是研究的热点之一。除了 Hadoop 外,还有其他处理工具。

1.3.3 大数据对高校的影响

大数据时代下数字化校园建设的总体目标应适应高速大容量、综合化、数字化和个人化等信息技术发展的大趋势,立足于决策、师生的客观需求,改善和提升学校的教学环境、科研环境、管理环境、生活环境、服务环境和人才培养环境,为实现学校教学质量、管理水平和办学效益的进一步提高提供技术支撑。科学合理利用学校整体数据,满足各部门需求,为师生们提供数据服务,为学校领导提供全局性的数据统计分析和决策支持。

1.4 高校大数据研究展望

大数据时代下的数字化校园建设最终落脚点和归宿,就是为学校的各项决策提供强有力的信息支持,使决策更加合理化、规范化和科学化。因此,高校应该紧跟时代步伐,利用大数据技术对校园信息化的数据进行大规模抓

取，并实施有效的深度数据分析，进而为教育资源优化、人才质量提高等方面提供科学的决策依据。目前，大多高校的大数据技术研究与应用尚未展开，只有少数高校开始着手研究。基于前期的探索，本节主要在高校大数据技术平台搭建与应用方面进行初步说明。

1.4.1 高校大数据应用

高校管理信息系统中汇聚着大量信息，从学生角度来看，包括个人基本信息、图书馆自修、食堂消费、住宿晚归等生活信息，选课、课后作业、借阅图书、成绩等学习信息，参与社团、竞赛、讲座等第二课堂信息；从教师角度来看，包含教学任务、课件等教学信息，论文著作、科研项目等科研成果信息；从管理者的角度来看，包含学校的师资信息、科研信息、教学信息、财务信息、资产信息等。同时，随着移动互联网以及物联网等技术的兴起，学校师生主动产生和由设备自动收集的信息越来越多，如微博、微信等社交信息，各类搜索点击记录信息等。上述信息存在着数据量大、结构复杂、产生频率快的特点，这些庞大、结构复杂的各种数据看似零散、无关联，但是在大数据时代，如果利用好这些数据，就能对高校的教学、科研、招生、就业起到很大的帮助，以下列举几点在高校中的应用。

1）实现个性化教育

以往的高校教学都是面向全班同学的整体化、统一教学，高校学生的基础和学习能力都不同，统一的教学安排势必会造成基础差、学习能力差的同学跟不上、听不懂、学不会，而学习能力强、基础扎实的学生则有可能被浪费了很多时间在已经学会的知识上。如果能针对不同的学生提供个性化的教育，那就能够大大提高学生的学习主动性、学习效率和高校的教学效果。在大数据时代，可以通过对学生在各个课程的课堂行为过程、做作业的行为过程、完成课程学习的情况、在学习中花费的时间等数据进行整合、分析，将不同学习特征的学生进行分类，从看似无关的大量数据中提炼能够真实地反映学生在教学过程中的个人的学习状态和特点，从而使教师能够在未来的教学过程中针对不同学生进行有个性化的教育，真正地去做到因材施教。

2）提高教学质量

如何提高教学质量一直是高校教师经常讨论、研究的问题。在大数据时