北京市高等教育精品教材立项项目

可下载教学资料

http://www.tup.tsinghua.edu.cn

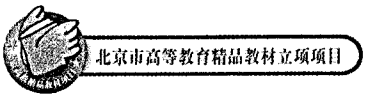
高等学校教材

计算机科学与技术

数据仓库与OLAP实践教程

何玉洁 张俊超 编著

清华大学出版社



高等学校教材

计算机科学与技术

数据仓库与OLAP实践教程

何玉洁 张俊超 编著

清华大学出版社

北京

内 容 简 介

数据仓库及联机分析处理是数据库技术发展和应用的一个新阶段，本书全面、详细地介绍了构建数据仓库以及进行多维数据分析的技术，并力求把数据仓库理论以及在该理论领域的相关应用尽可能完美地融合起来，其内容涵盖数据仓库的构建理论、构建示例、前端多维数据的展示及分析技术、对数据仓库及多维数据集的管理和维护等技术。本书以目前流行的 Microsoft SQL Server 2000 数据库管理系统作为实践平台，以便于读者实践。本书语言通俗易懂，实例丰富。

本书非常适合作为计算机、商科及相关专业本科学生学习数据仓库及多维数据分析技术的教材，同时也适合作为研究生数据仓库等课程的教材。

图书在版编目(CIP)数据

数据仓库与 OLAP 实践教程/何玉洁，张俊超编著. —北京：清华大学出版社，2008.3
(高等学校教材·计算机科学与技术)
ISBN 978-7-302-16810-2

Ⅰ. 数…　Ⅱ. ①何…　②张…　Ⅲ. 数据库系统—高等学校—教材　Ⅳ. TP311.13

中国版本图书馆 CIP 数据核字(2008)第 005359 号

责任编辑：魏江江　李玮琪
责任校对：白　蕾
责任印制：杨　艳
出版发行：清华大学出版社　　**地　　址**：北京清华大学学研大厦 A 座
http://www.tup.com.cn　　**邮　　编**：100084
c—service@tup.tsinghua.edu.cn
社 总 机：010-62770175　　**邮购热线**：010-62786544
投稿咨询：010-62772015　　**客户服务**：010-62776969
印 装 者：清华大学印刷厂
经　　销：全国新华书店
开　　本：185×260　**印　张**：22.5　**字　数**：543 千字
(附光盘 1 张)
版　　次：2008 年 3 月第 1 版　　**印　　次**：2008 年 3 月第1 次印刷
印　　数：1～4000
定　　价：37.00 元

出版说明

改革开放以来，特别是党的十五大以来，我国教育事业取得了举世瞩目的辉煌成就，高等教育实现了历史性的跨越，已由精英教育阶段进入国际公认的大众化教育阶段。在质量不断提高的基础上，高等教育规模取得如此快速的发展，创造了世界教育发展史上的奇迹。当前，教育工作既面临着千载难逢的良好机遇，同时也面临着前所未有的严峻挑战。社会不断增长的高等教育需求同教育供给特别是优质教育供给不足的矛盾，是现阶段教育发展面临的基本矛盾。

教育部一直十分重视高等教育质量工作。2001 年 8 月，教育部下发了《关于加强高等学校本科教学工作，提高教学质量的若干意见》，提出了十二条加强本科教学工作提高教学质量的措施和意见。2003 年 6 月和 2004 年 2 月，教育部分别下发了《关于启动高等学校教学质量与教学改革工程精品课程建设工作的通知》和《教育部实施精品课程建设提高高校教学质量和人才培养质量》文件，指出“高等学校教学质量和教学改革工程”是教育部正在制定的《2003—2007 年教育振兴行动计划》的重要组成部分，精品课程建设是“质量工程”的重要内容之一。教育部计划用五年时间（2003—2007 年）建设 1500 门国家级精品课程，利用现代化的教育信息技术手段将精品课程的相关内容上网并免费开放，以实现优质教学资源共享，提高高等学校教学质量和人才培养质量。

为了深入贯彻落实教育部《关于加强高等学校本科教学工作，提高教学质量的若干意见》精神，紧密配合教育部已经启动的“高等学校教学质量与教学改革工程精品课程建设工作”，在有关专家、教授的倡议和有关部门的大力支持下，我们组织并成立了“清华大学出版社教材编审委员会”（以下简称“编委会”），旨在配合教育部制定精品课程教材的出版规划，讨论并实施精品课程教材的编写与出版工作。“编委会”成员皆来自全国各类高等学校教学与科研第一线的骨干教师，其中许多教师为各校相关院、系主管教学的院长或系主任。

按照教育部的要求，“编委会”一致认为，精品课程的建设工作从开始就要坚持高标准、严要求，处于一个比较高的起点上；精品课程教材应该能够反映各高校教学改革与课程建设的需要，要有特色风格、有创新性（新体系、新内容、新手段、新思路，教材的内容体系有较高的科学创新、技术创新和理念创新的含量）、先进性（对原有的学科体系有实质性的改革和发展、顺应并符合新世纪教学发展的规律、代表并引领课程发展的趋势和方向）、示范性（教材所体现的课程体系具有较广泛的辐射性和示范性）和一定的前瞻

性。教材由个人申报或各校推荐(通过所在高校的“编委会”成员推荐),经“编委会”认真评审,最后由清华大学出版社审定出版。

目前,针对计算机类和电子信息类相关专业成立了两个“编委会”,即“清华大学出版社计算机教材编审委员会”和“清华大学出版社电子信息教材编审委员会”。首批推出的特色精品教材包括:

(1) 高等学校教材·计算机应用——高等学校各类专业,特别是非计算机专业的计算机应用类教材。

(2) 高等学校教材·计算机科学与技术——高等学校计算机相关专业的教材。

(3) 高等学校教材·电子信息——高等学校电子信息相关专业的教材。

(4) 高等学校教材·软件工程——高等学校软件工程相关专业的教材。

(5) 高等学校教材·信息管理与信息系统。

(6) 高等学校教材·财经管理与计算机应用。

清华大学出版社经过20年的努力,在教材尤其是计算机和电子信息类专业教材出版方面树立了权威品牌,为我国的高等教育事业做出了重要贡献。清华版教材形成了技术准确、内容严谨的独特风格,这种风格将延续并反映在特色精品教材的建设中。

清华大学出版社教材编审委员会

E-mail: dingl@tup.tsinghua.edu.cn

前言

数据仓库的概念自提出以来就备受业界的关注，其迅猛的发展势头和强劲的市场潜力已经为世界各大计算机研究机构和IT企业所关注。目前国际市场上也有了较为成熟的数据仓库产品及解决方案，但国内对数据仓库领域技术的研究和应用起步比较晚，发展也不很成熟。随着国内大型企业规模的日趋壮大，管理的日益先进，势必导致企业信息化和智能化的强烈需求，部署数据仓库解决方案成为一种必然的选择。而现今国内企业在这方面的应用仅限于报表阶段，数据的多维分析、数据挖掘和知识发现等更高级的数据处理和决策支持技术还没能普及开来。因此，在企事业单位和高等学校中讲授数据仓库领域的理论知识和应用技术就显得十分必要了，这也是本书写作的初衷所在。

本书着眼于理论与实践的结合，重点探讨了数据仓库和OLAP领域技术的应用层面。全书以数据仓库的构建流程作为整体线索，从源数据的清理转换到数据仓库的设计和实施，然后构建多维数据集，最后实施OLAP前端分析，书中详细介绍了各个环节所涉及的重要概念及应用技术。《数据仓库与OLAP实践教程》一书区别于同类书籍的特点有以下几个方面：

第一，实际案例驱动，面向实践与应用。数据仓库与OLAP实践是一项应用性很强的技术，很多这方面的书籍都借助Microsoft SQL Server 2000提供的示例数据库(如Northwind)来讲解这方面的应用，但这与现实的应用情况有很大的区别。示例数据库都是相对完整和规范的，而现实中的原始数据库则是多种多样的，存在着噪音数据、空值数据和类型不匹配的数据等问题，因此进行完整的清理和转换是一项重要且复杂的工作。作者在本书中使用了现实的案例，包括来自银行、烟草公司等多种实际数据，根据不同行业的分析需求分别介绍相关的技术，这不仅有益于读者了解应用的实际情况，同时还能帮助读者熟悉实际经济活动的相关背景，这也是本书独树一帜的一点。

第二，可操作性强，语言通俗易懂。数据仓库与OLAP实践实际上是一门综合学科，它涉及了信息检索、数据库、人工智能、模糊数学，甚至商科等各方面的知识，应用于现代企业经济活动的数据分析和决策支持方面，这也是国外很多著名大学商科学院也开设此类课程的原因所在。很多这方面的书籍偏重于数据仓库的理论、算法

等内容，这对很多该技术的使用者和学习者来说都是晦涩难懂的。而本书在实际案例的基础上，穿插了丰富的图示和操作步骤，读者和用户能够利用随书光盘中的数据，亲自实际操作和实施书中的案例。这对初次涉及该领域的学习者和急于应用该技术的用户都是十分有利的，另外对这方面技术的普及也颇具益处。

第三，技术介绍全面，应用层面广泛。本书融入了作者丰富的实践经验，在数据仓库与OLAP实践的各环节的介绍中，作者较全面地考虑了多种应用需求以及相应的技术。例如，在本书第7章《数据准备》中，作者根据自身实际经历的各种数据清理转换问题，创新地提出和总结了一整套数据清理和转换的技术方案，读者可以获得这方面的全面指导。此外，对于OLAP的前端分析策略的介绍，本书也提供了多种解决方案，其中Crystal Analysis的应用也是其他同类书籍所没有的。各种不同的前端分析方案面对着不同的分析需求，读者和用户可以根据自身遇到的实际情况来选择适合的方案。

本书前3章讲述了数据仓库与OLAP领域的主要理论和概念，起到提纲挈领的作用；第4章对OLAP主要工具的介绍可以让读者准备好自己的计算机；第5章是一个完整的金融案例，让读者对数据仓库和OLAP的实际应用有一个整体的了解；第6章关于设计数据仓库和OLAP策略的介绍让读者更深入理解数据仓库的实施过程和技术方法；第7章讲述了数据准备阶段的实用技术；而第8章DTS的应用则是将数据传输和装载至数据仓库的绝佳方案；第9章细致地讨论了有关维度构建和管理的话题；第10章则是面向多维数据分析的高级应用，读者能结合实际的例子学习应用计算成员、计算单元和对策等高级技术；第11章OLAP的前端分析策略选取了现今常见的几种前端分析工具，同样也是基于实际案例的应用；第12章介绍了微软Analysis Services专用的多维查询语言——MDX，为专业的IT人士对多维数据的编程和操作提供了接口和工具；第13章和第14章则总结了多维数据集和数据仓库的管理和优化；第15章作为本书的最后一章，对数据仓库和商务智能领域的现况和前景作了描述，并概要地介绍了数据挖掘技术的定义和概念，为进一步学习该方面的知识作了铺垫。

本书由何玉洁负责提出编写大纲，并进行全书的统稿及最终的审定，主要由张俊超负责执笔，同时张宏旭、于小倩、范洁参加了本书部分章节的编写工作，在此对他们的勤奋工作表示感谢。此外，本书还受到了国家审计署京津冀特派办的大力支持和帮助，作者从中获得了十分宝贵的实践经验和模拟数据，在此表示衷心的感谢！由于数据仓库和多维数据分析技术在国内的应用还处于探索阶段，更加之作者的水平和经验有限，有些方面研究得还不够深入和透彻，难免有错讹之处，还有待于读者和实践的检验，真诚希望广大读者批评指正！

何玉洁

2007年8月于北京

目 录

第1章

数据仓库与OLAP概述

人们在日常生活中经常会遇到这样的情况：超市的经营者希望将经常被同时购买的商品放在一起，以增加销售；保险公司想知道购买保险的客户一般具有哪些特征；医学研究人员希望从已有的成千上万份病历中找出患某种疾病的病人的共同特征，从而为治愈这种疾病提供一些帮助……对于以上问题，现有信息管理系统中的数据分析工具很难解答。因为无论是查询、统计还是报表，其处理方式都是对数据按指定的要求进行查询和处理，这种方式很难从这些数据中获取数据所包含的内在信息。随着信息管理系统的广泛应用和数据量的激增，人们希望能够提供更高层次的数据分析功能，为此，数据仓库应运而生。

数据仓库(Data Warehouse)已经进入到一个快速发展阶段。当现代企业开始重视信息的价值时，数据仓库就成为一个必然的选择。经济的发展和业务环境的变化是数据仓库发展的主要原因。20世纪90年代是全球以及中国经济急速发展的十年，激烈的竞争、企业间频繁的兼并重组，使得企业对信息的需求大大加强，这是数据仓库得以发展的根本原因。信息的潜在价值正在得到越来越多的关注，企业已经认识到充分利用信息是应对挑战的关键一步。数据仓库因而成为IT领域中最被关注的热点之一。

作为决策支持系统(Decision-making Support System，DSS)的辅助工具，数据仓库系统包含后台的数据仓库服务和前端的分析服务。在前端分析服务中，多维数据分析是现今最常用的分析技术之一，它又被称作联机分析处理(Online Analytical Processing，OLAP)。联机分析处理使得分析者、管理者和执行者可以快速、一致和互动地从多个角度访问数据，从而能发现更深层的信息。OLAP将原始数据转换成有用的信息，因此它能反映出影响企业商务活动的真正因素。除OLAP之外，数据挖掘(Data Mining)也作为强大的前端分析技术，逐渐在商业智能(Business Intelligence，BI)领域崭露头角。通过数据挖掘，管理者可以发现商业活动中那些隐秘的事实。数据挖掘使表面上没有什么关联的商业元素联系起来。

1.1 数据仓库

从20世纪80年代以来，联机事务处理(Online Transaction Processing，OLTP)数据库在实际应用的很多方面都发挥了重要的作用。随着历史数据的长期积累，众多企业的高层都意识到有必要利用这些积累数据进行分析，给下一步计划和决策提供参考。

数据仓库的目的就是合并和组织这些历史数据，以便对其进行分析并用来支持业务决

策。数据仓库中包含的历史数据通常是从各种不同的来源收集的，例如 OLTP 系统、文本文件或电子表格等。数据仓库整合这些数据，对其进行清理和转换，使其准确一致，并在此基础上进行组织，使其便于轻松高效地查询、分析。

1.1.1 数据仓库的概念和特点

数据仓库概念始于 20 世纪 80 年代中期，首次出现是在被称作“数据仓库之父”的 William H. Inmon 的“Building the Data Warehouse”一书中。随着人们对大型数据库系统研究、管理、维护等方面的认识的深入和不断完善，在总结、丰富、集中多年企业信息的经验之后，对数据仓库给出了更为精确的定义，即“数据仓库是在企业管理和决策中面向主题的、集成的、与时间相关的、不可修改的数据集合”。

根据上述数据仓库的定义，可知其有以下几个主要特点：

1. 面向主题

数据仓库围绕一些主题，排除对于决策无用的数据，提供特定主题的简明视图。主题是在较高层次上将企业信息系统中的数据综合、归类并进行分析和抽象，是对应企业中某一宏观分析领域所涉及的分析对象，是针对某一决策问题而设置的。面向主题的数据组织方式就是完整、统一地刻画各个分析对象所涉及的企业的各项数据以及数据之间的联系。

目前，数据仓库的实现主要是基于关系数据库，每个主题由一组相关的关系表或逻辑视图来具体实现。

2. 数据的集成性

数据仓库中存储的数据是从原来分散在各个子系统中的数据提取出来的，但并不是原有数据的简单复制，而是经过统一和整合的。其原因有以下两点：

(1) 原始数据一般不适合用于分析处理，因为原始数据的结构是面向事务处理设计的，因此，在进入数据仓库之前必须经过综合、计算，抛弃分析处理不需要的数据项，增加一些可能涉及的外部数据。

(2) 数据仓库中的每个主题所对应的原始数据在原分散的数据库中很可能存在重复或不一致，因而必须对数据进行统一，消除不一致和错误的地方，以保证数据的质量。否则，对不准确甚至不正确的数据分析得出的结果将不能用于指导企业作出合理和科学的决策。

对原始数据的集成是构建数据仓库中最关键、最复杂的一步，主要包括编码转换、度量单位转换和字段转换等。为了更好地支持对数据的分析，一般还需要对数据结构进行重组以及适当增加一些数据冗余。

3. 数据不可修改

从数据的使用方式上来看，数据仓库中的数据是不可更新的。这是指当数据被存放到数据仓库中之后，最终用户只能通过分析工具进行查询、分析，而不能修改其中存储的数据。

4. 数据与时间相关

数据仓库的数据不可更新，但这并不是说数据从进入数据仓库以后就永远不变。从数

据的内容上看，数据仓库中存储的是企业当前的和历史的数据。因而每隔一段固定的时间间隔后，需要再对源数据库中的数据进行抽取和转换，并集成到数据仓库中去。这就是说数据仓库中的数据随时间变化而定期地被更新，从而保证前端分析结论的时间有效性。

数据仓库中的结构信息、维护信息等均被保存在数据仓库的元数据中，数据仓库维护工作是由系统根据元数据中的定义自动进行的，或由系统管理员定期维护，最终用户不必关心数据仓库是如何被更新的细节。

1.1.2　数据仓库与传统数据库的比较

传统的关系数据库遵循一致的关系型模型，其中的数据或记录以表格的形式存储，并且能够用统一的结构化查询语言(Structured Query Language，SQL)对数据进行操作，因此它的应用常被称作联机事务处理(Online Transaction Processing，OLTP)。其重点在于完成业务处理，及时响应客户请求。关系型数据库能够处理大量的数据，但不能将其简单地堆砌就直接作为数据仓库来使用。

数据仓库主要工作的对象是多维数据，因此又称为多维数据库。数据仓库的数据以数组的方式存储，既没有统一的规律可循，也没有统一的多维模型可循。就应用而言，数据仓库应该具备极强的查询能力。数据仓库中存储的信息既多又广，但由于完成的是联机分析处理(OLAP)，因此并不追求瞬时的响应时间，只要在有限的时间内给予响应即可。实际上，OLAP包含交互式的数据查询，并提供多种分析方法，例如下钻到底层的细节信息上。因此数据仓库中的信息，尽管是多维的，仍然可以用具体的表格表示。

数据仓库与传统的数据库相比，其最大的区别是它们存储的数据。传统的数据库系统(OLTP系统)中的数据称为操作型数据，其值是不断变化的。而数据仓库中的数据通常被称作决策支持数据，其值保持相对稳定。表1-1列出了OLTP系统中的操作型数据和数据仓库中的决策支持数据的主要区别。

表1-1　操作型数据和数据仓库中的决策支持数据的区别

操作型数据	决策支持数据
面向应用：数据服务于某个特定的商业过程或功能	面向主题：数据服务于某个特定的商业主题，例如客户信息等。它是非规范化数据
细节数据，例如包含了每笔交易的数据	对源数据进行摘要，或经过复杂的统计计算。例如一个月中交易收入和支出的总和
结构通常不变	结构是动态的，可根据需要增减
易变性(数据可改变)	非易变(数据一旦插入就不能改变)
事务驱动	分析驱动
一般按记录存取，所以每个特定过程只操作少量数据	一般以记录集存取，所以一个过程能处理大批数据，例如从过去几年数据中发现趋势
反映当前情况	反映历史情况
通常只作为一个整体管理	可以分区管理
系统性能至关重要，因为可能有大量用户同时访问	对性能要求较低，同时访问的用户较少