高等学校信息管理示范教材

数据仓库与数据挖掘技术原理及应用

姚家奕 编著

http://www.phei.com.cn

電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY

高等学校信息管理示范教材

数据仓库与数据挖掘技术原理及应用

姚家奕　编著

電子工業出版社
Publishing House of Electronics Industry
北京 • BEIJING

内 容 简 介

本书从逻辑层次上分为导论、原理、技术和实践四大部分，第 1 章和第 2 章是导论部分，首先介绍了数据仓库与数据挖掘的基本概念，然后从应用层面介绍了数据仓库与数据挖掘技术在多个热点行业的最新应用情况。第 3 章至第 9 章是原理部分，系统介绍了数据仓库、OLAP 和数据挖掘技术的基本原理，以及关联规则分析算法、聚类分析算法、分类分析算法和序列模式分析算法。第 10 章至第 14 章是技术部分，以微软 SQL Server 2000 为数据管理平台，系统介绍了 OLAP 分析功能、多维数据集设计、维度和指标的建立、MDX 语言的应用、多维数据集的优化、数据挖掘和管理技术。第 15 章是实践部分，主要介绍了数据仓库系统的开发方法，以一个实际的数据仓库系统开发项目为背景，详细介绍了该系统的体系结构设计和模型设计。

本书既可作为高等院校硕士研究生和本科生的教材和参考书，也可作为程序设计人员的参考书。

图书在版编目（CIP）数据

数据仓库与数据挖掘技术原理及应用 / 姚家奕编著. —北京：电子工业出版社，2009.8
高等学校信息管理示范教材
ISBN 978-7-121-09398-2

I. 数… II.姚… III.①数据库系统－高等学校－教材 ②数据采集－高等学校－教材 IV.TP311.13 TP274

中国版本图书馆 CIP 数据核字（2009）第 135214 号

策划编辑：刘宪兰
责任编辑：李光昊
印　　刷：北京东光印刷厂
装　　订：三河市鹏成印业有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　邮编　100036
开　　本：787×1092　1/16　印张：28　字数：645 千字
印　　次：2009 年 8 月第 1 次印刷
定　　价：43.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

服务热线：（010）88258888。

前　言

回顾数据仓库和数据挖掘的发展历程，我们发现，这两个概念和理论原本是从不同的层次和角度提出，并且是相互独立，各自发展的。数据仓库是为提高分析和决策的效率和有效性，按照决策支持的需要对相关数据进行重新组织、建立单独的分析处理环境而出现的一种数据存储和组织技术。数据挖掘从技术的角度来讲，就是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程。从商业应用角度看，数据挖掘是是对商业数据中的大量业务数据进行抽取、转化、分析和模式化处理，从中提取辅助商业决策的关键信息，即从数据中自动发现相关商业模式。

经过数十年的发展和进化，数据仓库和数据挖掘逐渐融合在一起，形成一种新的模式，即“DW（数据仓库）＋OLAP（联机分析处理）＋DM（数据挖掘）”，如今，无论是从应用层次，还是从理论和技术层次，我们已经不能把二者分开，数据仓库成为数据挖掘的基础和数据保障，数据挖掘也不再神秘，高不可攀，已经从专家学者们独享的象牙塔中走出，把它的应用扩展到各个领域。

笔者长期从事数据仓库和数据挖掘的理论教学和科研项目研究，本书的编写就是在这些教学积累和科研成果基础上完成的。在本书的编写中坚持理论、技术和实践充分融和的原则，贯穿整书每一章节的基本思想是以解决实际问题为目标，书中的理论公式和技术方法都力图用一个具体的示例加以解释和说明，最后以一个实际开发的数据仓库项目为案例，将系统实际开发中不同环节和阶段遇到的问题给出具体的解决方法，不仅加深了对知识的理解，更增强了数据仓库系统实际开发能力。同时本书还配备了相应的实验教材。

本书从逻辑层次上分为导论、原理、技术和实践四大部分，第 1 章和第 2 章是导论部分，首先介绍了数据仓库与数据挖掘的基本概念，然后从应用层面介绍了数据仓库与数据挖掘技术在多个热点行业的最新应用情况。第 3 章至第 9 章是原理部分，系统介绍了数据仓库、OLAP 和数据挖掘技术的基本原理，以及关联规则分析算法、聚类分析算法、分类分析算法和序列模式分析算法。第 10 至第 14 章是技术部分，以微软 SQL Server 2000 为数据管理平台，系统介绍了 OLAP 分析功能，多维数据集设计、维度和指标的建立、MDX 语言的应用、多维数据集的优化、数据挖掘和管理技术。第 15 章是实践部分，主要介绍了数据仓库系统的开发方法，以一个实际的数据仓库系统开发项目为背景，详细介绍了该系统的体系结构设计和模型设计。

本书既可作为高等院校硕士研究生和本科生的教材和参考书，也可作为程序设计人员的参考书。

本书由姚家奕主编。在本书的编写过程中得到了张润彤教授、关中良教授的大力支持和帮助，并为本书提出了很多建设性的意见，在此一并感谢！同时还要感谢的是为本书的资料收集和文档编辑投入大量时间的研究生王一清、孔淑慧、石丹丹、葛洵洵和徐进华。最后应该感谢的是为本书的出版给予巨大支持的电子工业出版社。

本书难免会有许多不足之处，恳请专家、同行和读者提出宝贵意见。

编 者

2009 年 5 月

反侵权盗版声明

目　录

第1章

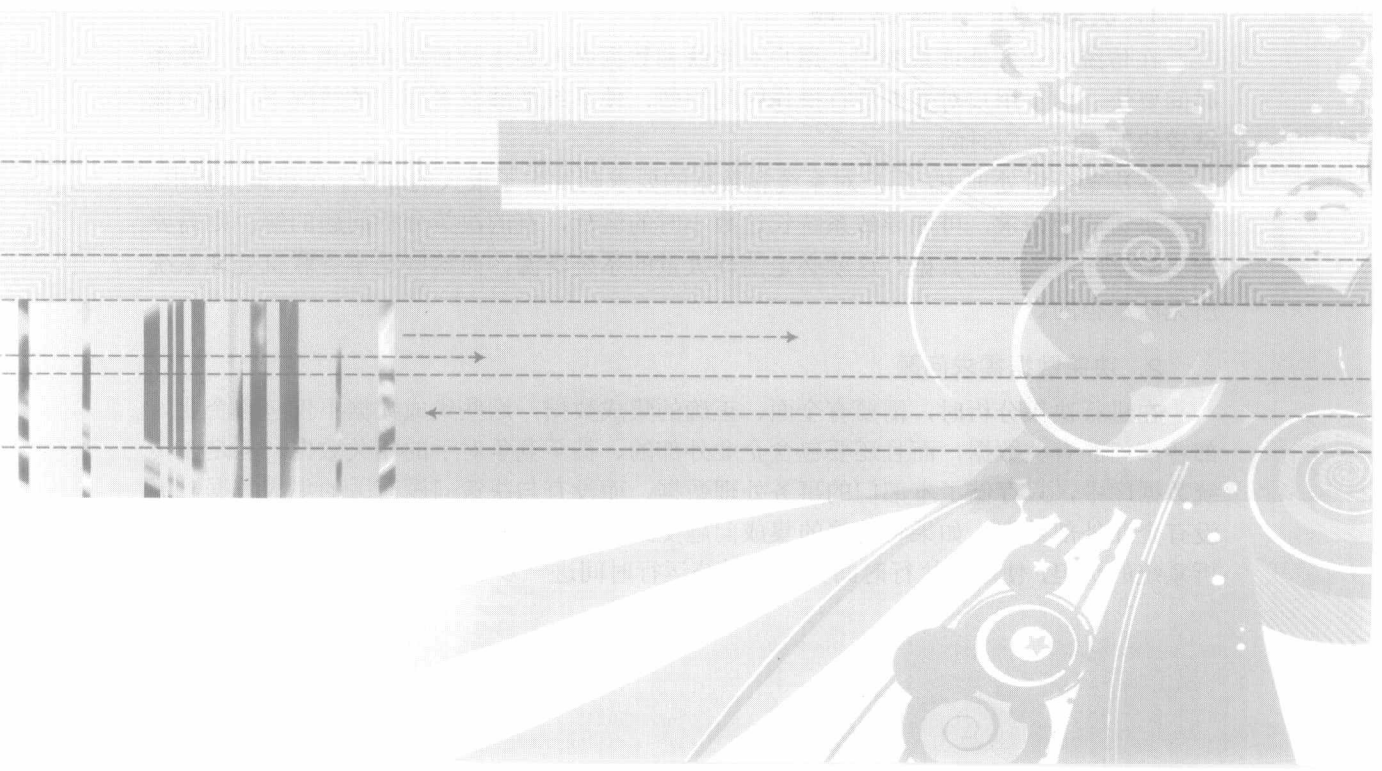
数据仓库与数据挖掘概述

本章引言

本章首先介绍了从数据库到数据仓库的演变过程，着重说明了传统数据库所面临的问题和局限性。要解决这些问题就必须在体系结构上加以变革，将操作型环境和分析型环境分离，使企业由原先以数据库为中心的生产环境过渡到以数据仓库为中心的生产环境。还介绍了数据挖掘的概念及其发展过程，并分析了数据挖掘与数据仓库的关系。描述了数据仓库和数据挖掘技术与相关学科和相关技术之间的关系，本章最后介绍了目前市场上数据仓库的主流产品。

本章重点

- 数据仓库产生的原因
- 数据仓库基本概念
- 数据仓库与数据库的区别
- 数据挖掘基本概念
- 数据挖掘与数据仓库的关系
- 数据仓库与数据挖掘的学科定位
- 对八大主流数据仓库产品的产品特点以及主要工具的了解



1.1　从数据库到数据仓库

信息技术的不断推广应用，将企业带入了一个信息爆炸的时代。每天都有潮水般的信息出现在管理者的面前，等待管理者去处理、去使用。这些管理信息的处理类型主要有事务型处理和分析型处理两大类。事务型处理，也就是通常所说的业务操作处理。这种操作处理主要是对管理信息进行日常的操作，对信息进行查询和修改，目的是满足组织特定的日常管理需要。在这类处理中，管理者关心的是信息能否得到快速的处理，信息的安全性能否得到保证，信息的完整性是否会遭到破坏。分析型处理则是指对信息做进一步的分析，为管理人员的决策提供支持。这种类型的信息处理在现代企业中应用越来越广泛，越来越引起管理人员的重视。管理信息的分析型处理，必须访问大量的历史数据才能完成，而不像事务型处理那样，只对当前的信息感兴趣。因此，在分析型处理中，产生了与操作型处理所采用的传统数据库有很大差异的数据环境要求。

传统数据库在操作型处理中获得了较大的成功，但是对管理人员的决策分析要求却无法满足。为实现管理人员的决策分析需要，在数据库基础上产生了能够满足决策分析所需要的数据环境——数据仓库（DW，Data Warehouse）。

1.1.1　数据库遇到的困境

传统数据库处理日常事务十分理想，但是要基于事务处理的数据库来帮助决策分析，就产生了很大的困难。其原因主要是传统数据库的处理方式和决策分析中的数据需求不相称，导致传统数据库无法支持决策分析活动。这些不相称性主要体现在决策处理中的系统响应问题、决策数据需求问题、决策数据操作问题和“蜘蛛网”问题。

1．决策处理的系统响应问题

在传统的业务处理系统中，用户对系统和数据库的要求是数据存取频率要高，操作时间要快。由于用户对数据进行操作时间短暂，使系统能够在多用户的情况下，也可保持较快的系统响应时间。

在决策分析处理中，用户对系统和数据的要求则发生了很大的变化。在决策分析中，有的决策问题请求，可能导致系统长达数小时的运行，有的决策分析问题的解决则需要遍历数据库中大部分数据。这就必定消耗大量的系统资源，而这些是事务联机处理系统所无法承担的。

2．决策数据需求问题

在进行决策分析时，需要有全面、正确的集成数据，这些集成数据不仅包含企业内部各部门的有关数据，而且还要包含企业外部的、甚至竞争对手的相关数据。但是在传统数据库中，只存储了本部门的事务处理数据，而没有与决策问题有关的集成数据，更没有企业外部数据。如果将数据的集成问题交给决策分析程序解决，将大大增加决策分析系统的负担，使原先执行时间冗长的系统运行时间进一步延长，用户将更加难以接受。