Broadview
www.broadview.com.cn

数据仓库经典图书

Enterprise Data Warehouse

企业级数据仓库（EDW）原理、设计与实践

王彦龙 编著

大型企业级数据仓库构建者的释疑库

原理——系统介绍、通俗易懂

设计——一线经验、操作性强

实践——多年累积、详尽叙述

電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

内 容 简 介

本书作者曾主持建成了我国证券市场惟一涵盖全市场业务数据的大型企业级数据仓库，他将其基于日常工作实践对数据仓库的概念和理论的理解及建设经验融入本书，详细地讲述了企业级数据仓库（EDW）的基本概念、规划、设计与实现，以及解决方案，并且还详细地描述了电信业、证券业和银行业数据仓库的案例，供数据仓库建设者借鉴。本书既是一本系统介绍数据仓库技术的通俗读物，又是一本数据仓库建设的实践指南，从本身架构到技术描述，从具体内容到实际操作，都不失为一本理论基础牢固、操作性极强的数据仓库经典图书。

本书可供关注、从事数据仓库的技术人员、管理决策人员参考阅读，也适合作为大中院校研究生的参考教材。

图书在版编目（CIP）数据

企业级数据仓库（EDW）原理、设计与实践 / 王彦龙编著. —北京：电子工业出版社，2006.9
ISBN 7-121-03109-4

I. 企… II. 王… III. 企业管理－数据库管理系统 IV. F270.7

中国版本图书馆 CIP 数据核字（2006）第 099545 号

责任编辑：孙学瑛
印　　刷：北京智力达印刷有限公司
装　　订：北京中新伟业印刷有限公司
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：787×980　1/16　印张：23.75　字数：431 千字
印　　次：2006 年 9 月第 1 次印刷
印　　数：5000 册　　定价：49.00 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，联系电话：（010）68279077；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。

序

Preface

我们处在一个信息爆炸的时代。二十多年前，我们难以掌握希望得到的信息，是因为我们拥有的数据太少。而今天，我们依然难以掌握希望得到的信息，却是因为我们拥有的数据太多。一方面我们缺乏足够的信息来支持科学的决策，而另一方面，多年积累的丰富数据却没有发挥作用。

我们在日常工作中希望将拥有的大量业务数据用于统计和分析，以帮助我们做出正确的决策。这个想法虽然简单而自然，但是在实际操作中我们却发现事情往往并不那么容易。丰富的业务数据往往被存放于分散的异构环境中，很难进行统一的查询访问；而且这些业务数据是按照日常业务处理流程来组织的，并不方便根据决策的主题进行汇总和分析。

数据仓库技术正是为了解决这一问题而产生的，它提供了一种进行分析处理的数据环境，从而使一个企业的信息化建设从支持日常业务操作上升到支持管理层的分析决策。在国外已有不少数据仓库建设的成功案例，它们使企业获得了可喜的效益。

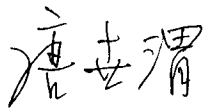
目前，国内数据仓库的建设方兴未艾。随着我国加入WTO，银行、电信、保险、证券等传统的垄断行业不得不面对国内外市场的激烈竞争。各个企业为了提高自己的市场竞争力，需要对自身企业的经营状况及整个市场相关行业的态势进行深入分析，进而做出有利的决策。这就需要建设自己的数据仓库系统。在这种背景下，《企业级数据仓库（EDW）原理、设计与实践》一书的出版适逢其时。

《企业级数据仓库（EDW）原理、设计与实践》一书的作者长期参与证券行业技术系统的一线建设工作和管理工作，直接参与了我国最早的证券交易联机事务处理系统的建设工作，主持建成了我国证券市场惟一涵盖全市场业务数据的大型企业级数据仓库，

在数据仓库系统的研发方面具有丰富的实践经验。

作者在书中将其基于日常工作实践对数据仓库的概念和理论的理解进行了系统的总结，对数据仓库涉及的关键技术作了深入浅出的介绍，对数据仓库建设中可能遇到的实际问题进行了深入的分析，并给出了切实可行的解决方案。同时，作者给出了多个行业的数据仓库典型案例，供建设者借鉴。

可以说，《企业级数据仓库（EDW）原理、设计与实践》一书既是一本系统介绍数据仓库技术的通俗读物，又是一本数据仓库建设的实践指南。此书的出版，对我国企业的大型数据仓库建设将不无裨益。



2006 年 8 月

唐世渭教授简介

唐世渭教授，1964 年毕业于北京大学数学力学系计算数学专业，毕业后留校任教至今。现为北京大学信息科学技术学院教授，博士生导师，中国计算机学会数据库专业委员会副主任。曾任北京大学信息科学中心主任、视觉与听觉信息处理国家重点实验室主任。多年来承担并完成“973”、“863”、国家科技攻关、国家自然科学基金等多项国家重点科研项目。曾获国家科技进步二等奖、三等奖各一项，省部级科技进步奖多项。在国内外杂志及会议上发表论文百余篇，著译作多部。目前主要研究方向为数据库系统、数据仓库和数据挖掘、Web 环境下的信息集成与共享、典型应用领域的信息系统等。

前 言

Introduction

在过去的几十年里，我们生成、收集和存储数据的能力不断提高。人们被淹没在数据的海洋中，却很难从繁杂的数据中获得决策的依据。我们需要一种有效的技术，帮助我们对数据库中的数据进行再加工，形成一个综合的、面向分析的环境，从而更好地支持决策制定。这种有效的技术就是数据仓库技术。

数据仓库是分析型数据库，是与操作系统相分离的、基于标准企业模型集成的、带有时间属性的、面向主题及不可更新的数据集合。它作为分析处理服务的基础，为制定决策提供所需的信息。

传统的数据库系统主要以面向事务处理为主的联机事务处理应用，无法满足决策制定时的分析处理要求。事务处理和分析处理在本质上存在很大差异，这种差异也导致了它们对数据有着不同的要求。数据仓库之父——W.H. Inmon 在其“Building the Data Warehouse”一书中，指出数据仓库中的数据应具备以下四个基本特征：

- 数据仓库的数据是面向主题的（subject-oriented）
- 数据仓库的数据是集成的（integrated）
- 数据仓库的数据是非易失的（nonvolatile）
- 数据仓库的数据是随时间不断变化的（time-variant）

数据仓库技术近几年得到了迅猛发展。随着市场竞争的加剧，尤其是我国加入WTO以后国外竞争者的加入，各个企业为了提高自己的市场竞争力，需要对自身企业的经营状况及整个市场相关行业的态势进行深入分析，进而做出有利的决策。在此过程中，很多大型企业纷纷开始建设自己的数据仓库系统。

作者长期参与证券行业技术系统的一线建设工作和管理工作，直接参与了我国最早的证券交易联机事务处理系统的建设工作，主持建成了我国证券市场惟一涵盖全市场业务数据的大型企业级数据仓库。在此，作者很高兴能有机会把在数据仓库系统研发和建设方面积累的实践经验进行总结并与大家分享，以期对数据仓库的建设者们有所帮助。

本书系统地总结了作者基于日常工作实践对数据仓库概念和理论的理解，对数据仓库涉及的关键技术做了深入浅出的介绍，深入地分析数据仓库建设中可能遇到的实际问题，并给出了切实可行的解决方案，同时，书中还给出了多个行业的数据仓库典型案例供大家参考。

本书是针对数据仓库的开发者、管理者、设计者以及数据仓库建设的其他相关人员而写的，作者真诚地希望读者在数据仓库实践中遇到的大多数问题都能在本书中找到答案。由于本书对数据仓库的概念和理论都做了系统的介绍，因此对于数据仓库的初学者而言，也是一本很好的入门读物。

作者简介

王彦龙 工学硕士、经济学博士、全国金融标准化技术委员会证券分技术委员会副主任委员。曾在深圳证券交易所、中国证监会工作多年，历任深圳证券交易所电脑工程部经理、深圳巨潮信息公司总经理、中国证监会市场监管部交易监管处处长，现任中国证券登记结算公司总工程师。作者对中国证券交易及登记结算技术系统建设具有实际的领导经验，曾主持并参与深圳证券交易所交易系统建设、中国证券登记结算领域多项技术系统的建设，主持搭建了国内首家证券类互联网信息系统，并主持搭建了涵盖沪、深证券市场的统一的 TB 级数据仓库。

目 录

Contents

第 1 章　数据仓库的基本概念

本章向读者介绍关于数据仓库的基本概念。首先介绍什么是数据仓库，然后将数据仓库和 OLTP 数据库、数据集市进行比较，阐述它们之间的不同。随着数据仓库技术的不断发展，许多新技术涌现了出来，在数据仓库技术的发展一节将对此进行详细描述。最后，向读者描绘数据仓库的投资回报，IDC 公司的调查结果表明，数据仓库的确为企业提供了巨大的收益。

1.1　什么是数据仓库

社会的需求极大地推动了技术的发展。数据库技术最初产生于 20 世纪 60 年代中期，从最初的网状、层次数据库，到日益成熟并广泛应用的关系数据库，大量的数据已经在数据库系统中积累起来。然而，数据的丰富却未能很好地解决知识贫乏的问题，人们不断尝试对数据库中的数据进行再加工，形成一个综合的、面向分析的环境，从而更好地支持决策制定。这些尝试促使数据仓库的思想逐渐形成。但对于什么是数据仓库，许多人提出了不同的定义。

“数据仓库是分析型数据库，作为分析处理服务的基础，用来存放海量的只读数据，为制定决策提供所需的信息。”

“数据仓库是与操作型系统相分离的、基于标准企业模型集成的、带有时间属性的（即与企业定义的时间区段相关）、面向主题及不可更新的数据集合。”