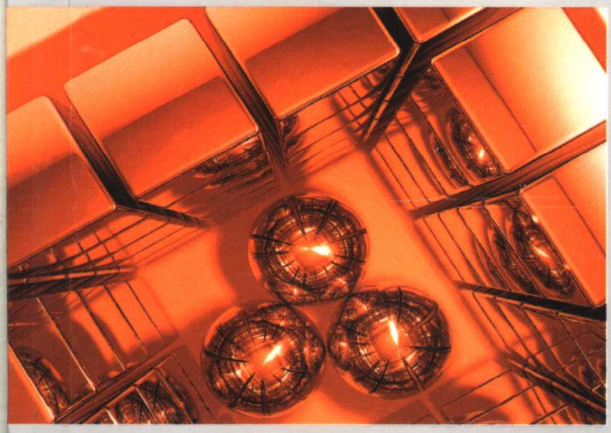
Broadview®
www.broadview.com.cn

数据仓库与数据挖掘技术应用丛书



Data Warehousing Architecture Design and Development: Build Information System Pyramid

数据仓库结构设计与实施

——建造信息系统的金字塔

池太崴　编著

数据仓库建立与开发的实践经验之理论总结

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

数据仓库与数据挖掘技术应用丛书

数据仓库结构设计与实施

——建造信息系统的金字塔

Data Warehousing Architecture Design and Development:
Build Information System Pyramid

池太崴　编著

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

内 容 简 介

全书主要描述了数据仓库设计与开发生命周期的各个阶段及其相应的技术结构框架。另外，还提供了如支票信用认可网络系统、某飞机制造公司单源生产数据、汽车销售管理网络系统等多个工程开发的成功案例。同时，本书还介绍了数据仓库开发的过程和策略，主要有跳跃（蛙跳）式发展、数据仓库系统多层次的结构——内核与外壳、数据仓库应用开发的要点与特征、数据仓库设计、质量保障、资源、团队、技能等内容。最后，通过一系列程序和实例讲解了数据仓库的建立过程，以及 SQL 程序在 Oracle 9i 和 SQL Server 分析服务器上的实施。

全书以成功实践为基础，理论与技术实践密切结合，可作为高等院校信息技术和管理专业、数据库专业教学与研究的教材，同时也适合从事信息系统研究与工程应用开发的广大科技人员作为参考读物。

图书在版编目（CIP）数据

数据仓库结构设计与实施：建造信息系统的金字塔 / 池太崴编著. —北京：电子工业出版社，2005.11
（数据仓库与数据挖掘技术应用丛书）
ISBN 7-121-01861-6

I. 数… II. 池… III. 数据库管理系统—结构设计 IV. TP311.138

中国版本图书馆 CIP 数据核字（2005）第 122557 号

责任编辑：孙学瑛
印　　刷：北京智力达印刷有限公司
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
经　　销：各地新华书店
开　　本：787×980　1/16　印张：21　字数：337 千字
印　　次：2005 年 11 月第 1 次印刷
印　　数：4000 册　定价：45.00 元

凡购买电子工业出版社的图书，如有缺损问题，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。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。

前　言

数据仓库是近 10 年来兴起的一种新的数据库应用，越来越被广泛地用做决策支持系统的基础。各大数据库厂商纷纷宣布支持数据仓库，并提出一整套用于建立和使用数据仓库的产品。同时，国际上许多重要的学术会议，如超大型数据库国际会议（VLDB）、数据工程（Data Engineering）国际会议等，都出现了专门研究数据仓库（Data Warehousing，简称 DW）、联机分析处理（On-Line Analytical Processing，简称 OLAP）、数据挖掘（Data Mining，简称 DM）的组织。

数据仓库的新概念、新方法、新技术在信息技术领域已成为研究和应用的热点，并日益成熟，是信息技术应用领域最前沿的技术。实践证明，数据仓库在提高决策支持水平、信息质量、应变能力等方面具有重要意义。它在操作型数据库的基础上对数据的进一步集成和分析提出了更明确的目标和解决方案，这对于信息技术领域发展中的全面而长远的规划尤其重要，能够加快信息技术实施速度并少走弯路，避免时间、人力、资源的浪费及重复建设。另外，数据仓库这一信息技术在理念、概念、方法、应用技术、定义功能等方面得到了广泛应用，给用户带来了巨大的竞争优势。对我国各领域和许多企业而言，在建立或发展自己的信息系统的过程中，如何在现有的数据库上建立数据仓库，怎样建立数据仓库，如何使数据仓库真正成为战略决策的基础系统等问题显得日益突出。

本书将从数据仓库技术背景、技术结构框架、开发和应用等方面，结合作者在数据仓库技术实施过程中的实际经验，深刻阐述数据仓库开发在各个阶段的特点和策略运用，以及在管理信息系统中的位置和作用。

数据仓库是管理信息系统的“上层建筑”，它集成了许多不同的源数据系统，从而构成中央式的信息平台，并实现了对管理决策分析的支持。今天，越来越多的部门和机构开始接受并开发数据仓库，并把它作为信息集成的解决方案和决策支持系统工具，以迎接当今日益发展的信息化的挑战。

笔者于 20 世纪 90 年代初在美国麦克尼斯州立大学计算机系获得硕士学位，在数据仓库、数据库设计和系统分析与集成方面有着丰富的理论知识和实践经验。特别是在过去的十几年中，几乎一直战斗在数据仓库建立和开发的前沿，在北美成功地完成了多个大型应用项目，有丰富的总体系统规划实践及大型项目的运作经验，并指导过许多项目的设计与开发，取得了显著的成效，博得了用户的好评。为了把这些成功的实践经验介绍给国内的广大读者和应用开发研究人员，作者在百忙中投入了大量的精力和时间，完成了本书。实践经验的理论总结是该书的最大特点。

本书旨在介绍和探索数据仓库的基本技术和结构概念（如多层次的结构），这种研究是非常重要的，因为基本技术和结构为将来的数据仓库开发，以及各种决策支持系统奠定了基础。当我们对许多信息系统开始进行以集成为目标的基础结构改造的时候，它将变得更为重要。希望读者能够从中获益，有所借鉴。

在本书的编写过程中，罗晓沛教授、刘德贵研究员，以及朱军、朱炜、池太峰、赵玉梅等人提供了大力的帮助和支持，在此深表谢意。

时间仓促，错漏之处在所难免，希望读者批评指正。

池太崴

目　录

第 1 章　数据仓库技术与应用概述 …… 1

1.1　数据仓库的基本概念 …… 2

1.1.1　数据仓库的系统体系 …… 2

1.1.2　数据仓库的应用目标 …… 2

1.2　数据仓库与常规事务处理数据库的区别与联系 …… 3

1.2.1　从数据仓库到操作型数据库——数据仓库的根与源 …… 3

1.2.2　数据仓库与传统数据库的区别 …… 3

1.3　数据仓库的产生原因 …… 5

1.3.1　数据囚笼现象 …… 5

1.3.2　信息孤岛现象 …… 5

1.3.3　相互矛盾的信息流 …… 6

1.3.4　集成的解决办法 …… 6

1.3.5　动力和动机 …… 7

1.4　管理信息系统的“上层建筑” …… 8

1.4.1　管理层次的概念 …… 8

1.4.2　中层和上层管理存在的系统真空 …… 9

1.4.3　数据仓库系统应用的基本作用 …… 9

1.4.4　数据仓库应用的基本目标 …… 11

1.4.5　数据仓库应用成功的保障 …… 11

1.5　电子商务与电子政务 …… 12

1.5.1　现代社会中的电子商务与电子政务 …… 12

1.5.2　以客户为中心的现代社会环境 …… 14

1.5.3　电子商务与电子政务提高了服务效率 …… 14

1.6　数据仓库的 ROI（回报投入比） …… 17

1.7　联机综合分析系统中数据仓库的应用 …… 18

1.8 挑战和趋势……20

第 2 章 数据仓库的总体结构……24

2.1 金字塔结构……24
2.2 数据仓库的结构与环境……28
2.3 准备区——数据源和数据仓库之间的过渡……31
2.4 元数据与模型……31
2.4.1 元数据的定义……33
2.4.2 元数据的作用……33
2.4.3 元数据和模型的整建……33
2.4.4 命名法……34
2.4.5 元数据存储区……35
2.4.6 元数据的维护和应用……36
2.4.7 元数据的定义和管理……37
2.4.8 统一元数据标准和元数据交换……42
2.5 多维数据结构……45
2.5.1 星型结构……45
2.5.2 雪花型结构……47
2.5.3 混杂型结构……49
2.5.4 度量应用举例……49
2.6 映像……49
2.6.1 映像的含义……51
2.6.2 数据迁移和转换的过程……51
2.6.3 抽象与映像层次……55
2.6.4 应变式映像策略……56
2.6.5 映像类型……57
2.7 滚动综合数据……58
2.8 联机分析处理……63
2.8.1 联机分析处理——数据仓库的自然延伸……63
2.8.2 联机分析处理系统的集成……64

2.8.3 维的作用 ····· 64
2.8.4 对多维数据方阵的链接和分析 ····· 65
2.8.5 方阵系列的设计要点 ····· 66
2.8.6 总计数据的自动更新 ····· 66
2.8.7 报表构架 ····· 66
2.8.8 联机分析处理（OLAP）的解决办法 ····· 67
2.8.9 表示工具 ····· 67
2.8.10 表示工具的预处理 ····· 67
2.9 数据发掘 ····· 68
2.9.1 数据发掘的重要性 ····· 69
2.9.2 数据发掘的方法与技术 ····· 69
2.10 实现闭环的联机分析处理 ····· 70
2.11 卸载操作型数据库与保护数据源 ····· 73
2.11.1 数据源——企业最重要的信息资产 ····· 73
2.11.2 操作型事务处理数据库的特征 ····· 73
2.11.3 决策支持数据库系统的特点 ····· 73
2.11.4 两种作业混合的弊端 ····· 74
2.11.5 回顾过去作业的局限性 ····· 74
2.11.6 卸载 ····· 75
2.11.7 双赢的解决办法 ····· 77
2.12 数据仓库的三要素 ····· 78
2.13 多维总计方阵 ····· 80
2.13.1 从基本数据到综合信息 ····· 80
2.13.2 方阵是联机分析的基础结构 ····· 80
2.13.3 方阵的类型 ····· 81
2.13.4 方阵的卸载与底层数据表的屏蔽 ····· 84
2.13.5 刷新 ····· 85
2.13.6 方阵的设计要点 ····· 86
2.13.7 从数据仓库基本数据（事实/维）到最终分析报告的映像 ····· 87
2.14 ETL（提取—转换—加载）从数据源到目标 ····· 88

2.14.1 数据的启程 ····· 89
2.14.2 数据标准化的准备工作和数据清洗的工具字典 ····· 89
2.14.3 粒度与聚合数据 ····· 89
2.14.4 魔力无边的巨型章鱼 ····· 90
2.14.5 数据仓库的数据追加 ····· 90
2.14.6 提取—转换—加载处理的映像过程 ····· 91
2.14.7 作业顺序、依赖关系和进程控制 ····· 91
2.14.8 从数据源进入数据仓库到以分析报表输出 ····· 92
2.14.9 数据提取—转换—加载的主要流程和会话期流程 ····· 93
2.15 从数据源到目标——Informatica ····· 96
2.16 数据仓库在因特网环境下的应用 ····· 98
2.16.1 客户-服务器系统的特点 ····· 99
2.16.2 因特网数据仓库的特点 ····· 99
2.16.3 设计指南 ····· 99
2.16.4 安全性技术 ····· 102
第 3 章 数据仓库应用实例 ····· 104
3.1 分布式数据仓库——独立的数据库接口 ····· 104
3.2 共享式数据仓库——共享式支票信用认可网络 ····· 107
3.3 某飞机制造公司——单源生产数据 ····· 108
3.4 汽车销售管理网络——数据仓库支持下的联机分析报表 ····· 109
第 4 章 数据仓库应用开发的策略与过程 ····· 111
4.1 数据仓库开发策略 ····· 111
4.2 跳跃（蛙跳）式发展 ····· 113
4.2.1 数据仓库的演变史 ····· 113
4.2.2 建立真正的数据仓库 ····· 114
4.3 数据仓库系统平台 ····· 116
4.3.1 观察数据仓库系统的基本结构 ····· 116
4.3.2 多层结构环境 ····· 116
4.3.3 多层次、多分区系统 ····· 117

4.3.4 坚实的胡桃 …… 118
4.3.5 表示层与内核的部署 …… 118
4.3.6 应用软件的基本结构 …… 119
4.4 数据仓库应用开发的要点与特征 …… 123
4.4.1 数据仓库应用的命题/主题确定 …… 123
4.4.2 往复循环式开发数据仓库 …… 128
4.4.3 建立数据集市 …… 133
4.5 数据仓库设计质量 …… 134
4.5.1 数据仓库质量的重要性 …… 134
4.5.2 数据质量保障 …… 135
4.5.3 数据质量保障的环境和各个处理环节 …… 135
4.5.4 错误检测 …… 137
4.5.5 质量保障系统 …… 137
4.5.6 及时发现错误 …… 138
4.5.7 错误追踪 …… 138
4.5.8 解决劣质数据 …… 144
4.6 数据仓库应用开发保障技术 …… 144
4.6.1 知识与知识产权的维护 …… 145
4.6.2 团队 …… 147
4.7 数据仓库安全性与有关技术 …… 152
4.7.1 识别安全威胁的类型与攻击方法 …… 152
4.7.2 安全性防范思想与布局 …… 154
4.7.3 安全性策略与技术 …… 155
4.7.4 数据仓库安全性的应用结构设计技术 …… 156
第 5 章 数据仓库设计与应用开发 …… 160
5.1 数据仓库的概念设计 …… 160
5.1.1 概念设计 …… 161
5.1.2 元数据定义及管理 …… 162
5.1.3 数据结构概图 …… 162

5.1.4 数据仓库的基本表……163
5.1.5 从逻辑设计到物理设计……164
5.2 数据仓库的物理设计……166
5.2.1 事实表设计……166
5.2.2 维数……170
5.2.3 分区……173
5.2.4 索引设计……175
5.2.5 完整性约束设计……176
5.2.6 实体化视图设计……177
5.3 数据提取—转换—加载（ETL）……188
5.3.1 建立事件映像……189
5.3.2 建立视图或实体化视图与视图模拟……189
5.3.3 ETL 过程举例……189
5.3.4 提取—转换—加载的方法……192
5.3.5 数据的标准化与规范化……193
5.3.6 数据清洗与实例……195
5.3.7 数据提取—转换—加载工具……198
5.3.8 数据提取……200
5.3.9 加载和转换……204
5.3.10 数据提取—转换—加载的主流程……209
5.4 综合管理……212
5.4.1 总体构架……212
5.4.2 汇总准备……214
5.4.3 报表准备工作的基础——从数据仓库生成并刷新实体化视图……214
5.4.4 刷新实体化视图……214
5.4.5 监控数据仓库的刷新……216
5.4.6 实体化视图的管理要点……218
5.5 联机分析处理（OLAP）……219
5.5.1 SQL 与综合函数……219
5.5.2 多维分析技术……220

5.5.3　数据仓库 SQL 总计分析语句结构与流程 ····· 221
5.5.4　综合 SQL 和函数的应用 ····· 223
5.5.5　SQL 和分析函数 ····· 236
5.6　报表发布 ····· 243
5.6.1　表示系统软件工具的联用——从后台到前台 ····· 243
5.6.2　建立报表的过程 ····· 245
5.6.3　对多维方阵的钻入/聚合操作 ····· 245
5.6.4　表示工具的预处理 ····· 246
5.6.5　应用 SQL 分析服务器 ····· 246
5.7　报表系统构架 ····· 249
5.7.1　报表系统构架及其支撑结构 ····· 250
5.7.2　从数据库生成 XML 数据 ····· 250
5.7.3　建立报表函数库 ····· 259
5.7.4　建立报表程序库 ····· 287
5.7.5　报表系统构架及其支撑结构 ····· 301
5.7.6　应用表函数 ····· 312
附录 A　技术词汇 ····· 319
参考文献 ····· 321

第 1 章　数据仓库技术与应用概述

乱生于治，治乱，数也。

——孙子

数据仓库是以关系数据库、并行处理与分布式处理技术，以及联机分析处理等技术的发展为基础，为解决当前企业和组织中虽然拥有大量数据但信息贫乏（难以利用）的现状而提出的，是一种对不同系统数据实现集成和共享的综合性解决方案。

从普通数据库与数据仓库的关系来看，人们把普通数据库技术称为传统的数据库技术。传统的数据库往往是以单一的数据资源（即以数据库为中心）进行事务处理、批处理、决策分析等各种数据处理工作。数据处理模式主要划分为两大类：操作型处理和分析型处理（或信息型处理）。操作型处理也叫事务处理，是指对数据库联机的日常操作，它通常是对一个或一组记录的查询和修改，主要是为企业的特定应用服务的，基本上满足了响应时间、数据的安全性和完整性的需要；分析型处理则用于管理人员的决策分析，往往是大规模的、批量的计算作业，经常要访问大量的历史数据。也就是说，传统数据库系统能够完成企业的日常事务处理工作，但很难达到实现数据分析处理的要求，也无法满足数据处理多样化的要求。随着用户需求的发展，操作型处理和分析型处理的分离就成为必然。

近年来，随着信息化的发展和技术的进步，信息已成为人类社会不可或缺的重要资源。社会的信息化使得信息量急剧增长。面对数据量的急剧增长和应用要求的不断提升，数据库技术的应用和发展也有了更高的作用和价值。数据库技术一直力图使自己能胜任当前的发展变化，完成从事务处理、批处理到分析处理的各种类型的信息处理任务。虽然业务扩充了，但还是要在统一数据格式、统一数据模型下来实现业务操作的数据处理。对于决策分析，在业务操作

层面上进行分析判断还存在着很大的局限性。于是，人们尝试对来自操作型处理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，以更好地支持决策分析，这就形成了数据仓库（Data Warehousing，简称 DW）技术。作为决策支持系统（Decision-making Support System，简称 DSS）的数据仓库系统包括：数据仓库技术、联机分析处理技术（On-Line Analytical Processing，简称 OLAP）、数据挖掘技术（Data Mining，简称 DM）。

数据仓库弥补了原有数据库的不足，将原来的以单一数据库为中心的数据环境发展为一种新的体系环境。它具有一种新的数据处理结构体系，能够将不同环境、不同系统的数据统一起来，以形成综合的中央数据仓库。

1.1 数据仓库的基本概念

业界公认的数据仓库概念创始人 W.H.Inmon 在《建立数据仓库》一书中对数据仓库给出的定义是：数据仓库就是面向主题的、集成的、稳定的（不可更新）、随时间变化（不同时间）的数据集合，它用以支持经营管理中的决策制定过程。

1.1.1 数据仓库的系统体系

数据仓库是以计算机应用为基础的信息系统，用来支持在各领域的决策分析。数据仓库作为一个集成了许多数据源的中央数据库系统，从许多不同的（分散的、互不联系的）联机事务处理数据源收集和提取数据，并通过一系列汇总计算将数据组织成易于分析的形式，从而为企业提供了一个信息集成平台，为管理人员和决策者迅速地提取信息并回答有关业务运作的问题提供支持。因此，数据仓库是企业信息资产的核心，是管理信息系统的“上层建筑”。

1.1.2 数据仓库的应用目标

数据仓库和普遍的事务处理数据库不同，它是面向主题（以主题为导向）

的，支持商务决策而不是事务处理。它拥有许多优化设计的层次、总计方阵系列和结构化的查询功能，并以总计/综合系统为构架。基于对数据快速和有效的分析，数据仓库可为决策系统提供强有力的支持。在开发人员和用户的协同配合和精心设计下，它能够实现对数据的一系列转化，包括从数据到信息，从信息到知识，最终到商业智能。

数据仓库最根本的特点之一是存放数据，而且这些数据包含历史数据，并且来源于各种数据库。数据仓库的建立并不是要取代操作性事务处理数据库（事务处理数据库在企业的信息环境中承担的是日常业务操作的任务），相反，它依赖于操作性事务处理数据库，并以此为基础，建立一个综合的和完善的信息分析应用系统，用于支持各级高管理层决策分析。数据仓库是数据库技术的一种新模式，一般也是用关系数据库系统来管理其中的数据。

1.2　数据仓库与常规事务处理数据库的区别与联系

首先让我们讨论数据仓库和普通的数据库之间的联系，然后再讨论数据仓库和普通的数据库有什么不同。

1.2.1　从数据仓库到操作型数据库——数据仓库的根与源

数据仓库的数据源来自操作型数据库，即联机事务处理系统。在数据提取—转换—加载处理系统的控制下，数据要经历“艰难”的历程，才能完成一系列的转换，变成对终端用户有用的信息，形成一个新的集成系统（联机分析系统），并用于决策分析。

1.2.2　数据仓库与传统数据库的区别

数据仓库区别于传统的数据库系统。对数据仓库而言，主要特点是集成和分析能力。表 1.1 是数据仓库与传统操作型数据库的比较。

表 1.1　数据仓库与传统数据库的比较

比较内容	数据仓库的特征	常规事务处理数据库
目标	OLAP 联机分析处理	OLTP 联机事务处理
作用	面向主题	面向过程
活动特征	分析式	事务处理
构成	集成	不同的、分散的
内容	不更改性	更改的
时间性	时序性、历史性	当前的
基础结构	多维型	关系型
关系结构	星型/雪花型结构或混杂型结构	3NF 三级范式
终端用户	多为管理人员和决策者	多为专业及操作人员

其中，雪花型结构是在星型结构上的进一步扩展。而对于混杂型（Hybrid）结构，由于兼顾多维星型结构和关系型范式结构的特点，因此具有某些杂交优势，它往往是根据实际应用情况实施结构优化设计的结果。

由于数据仓库依赖于来自操作型数据库的数据，因此操作型数据库是数据仓库的根与源，就像一棵大树依赖于它的根系存活一样。没有数据源，数据仓库就无从谈起；切断数据源，数据仓库就会“死掉”。图 1.1 可进一步说明它们之间的关系。

图 1.1　数据仓库与传统数据库的关系

1.3　数据仓库的产生原因

事务处理环境不适宜决策支持系统（Decision Support System，简称 DSS）应用的主要原因可以从以下 5 个方面来分析。

1.3.1　数据囚笼现象

一个人们常议论的问题是：已经有了数据库系统在处理我们的日常业务，为什么还要数据仓库？

可是，要用当前数据作为管理的决策支持，数据库系统能够胜任吗？

许多企业或组织机构在管理运作中，已经积累了大量的数据，包括业务运作、客户、产品和人员等。但是这些数据却被埋藏在计算机系统中未加以或难以利用，尤其是那些对于管理决策者有着重要意义的数据分析，没有被加以分析，就不能发挥其应有的作用和潜力。

如果企业或组织没有能力及时获得有价值的信息，就很难在今天这样迅速变化的环境中保持其竞争力。也就是说，存在着所谓的数据丰富、而信息贫乏的数据囚笼现象。

由于已经对计算机设备、数据库系统等进行了可观的投资，却对已产生的大量有价值的数据没有充分加以利用，这就是一个巨大的浪费。一些高级决策管理者曾不无感慨地说："我们在计算机系统上花了很多钱，它却不能回答我们的问题！"

1.3.2　信息孤岛现象

另一个问题是：很多部门都已经在使用计算机和数据库，为什么还要数据仓库？

可是，用当前系统进行组织机构、全企业、全行业的综合管理和宏观调控，能够得到决策支持吗？当前常规系统是否相互之间完全兼容和集成呢？

事实上，旧系统往往是在过去不同的时期被不同的开发者开发的，这些系统通常是根据某些特定的要求制作的，并且分布于不同的系统平台上，同时信