

Data Warehousing Fundamentals

数据仓库基础

[美] Paulraj Ponniah 著

段云峰 李剑威 韩洁 宋美娜 译

内 容 简 介

本书全面而详细地讲述了关于数据仓库每一个重要部分的内容,包括计划、需求、体系、基础、结构、设计、数据准备、信息传递、配置和维护等。本书编排合理,每章提供本章主题,本章小结,可以使读者将每一个概念和技术同数据仓库的实践和市场结合起来;还提供复习思考题和练习供读者巩固学习到的知识。总之,本书几乎涵盖了数据仓库领域所有方面的知识,如果你想成为数据仓库领域专家,它是值得一读的。

本书是专门为IT专业人员而量身定做的介绍数据仓库知识的书籍,适合于想掌握数据仓库基础知识的系统分析员、程序员、数据分析师、数据库管理员、项目经理和软件工程师阅读,还非常适合作为大中院校相关专业的教科书或培训用书,供自学人员、大学课程或科研机构使用。

John Wiley & Sons, Inc.

Copyright ©2001 by John Wiley & Sons, Inc.

All rights reserved. Authorized translation from the English language edition published by John Wiley & Sons, Inc.

本书简体中文专有翻译出版版权由John Wiley & Sons Inc. 授予电子工业出版社未经许可,不得以任何方式复制或抄袭本书的任何部分。

版权贸易合同登记号 图字: 01-2003-1540

图书在版编目(CIP)数据

数据仓库基础 / (美)波尼阿(Ponniah, P.)著;段云峰等译. —北京:电子工业出版社, 2004.4

(数据仓库与数据挖掘技术应用丛书)

书名原文: Data Warehousing Fundamentals

ISBN 7-5053-9784-2

I. 数... II. ①波...②段... III. 数据库系统 IV. TP311.13

中国版本图书馆 CIP 数据核字 (2004) 第 023051 号

责任编辑: 孙学瑛

印 刷: 北京市增富印刷有限公司

出版发行: 电子工业出版社 <http://www.phei.com.cn>

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: 37.5 字数: 643 千字

印 次: 2004 年 4 月第 1 次印刷

印 数: 4000 册 定价: 69.00 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 zltz@phei.com.cn, 盗版侵权举报请发邮件至 dbqq@phei.com.cn。

译者序

近年来，在数据仓库领域，陆续出现了很多从国外引进翻译过来的书籍，Ponniah 博士的《数据仓库基础》是其中比较出色的一本书。作者亲自做过有关数据仓库方面的教学工作，对学生到底需要学习哪方面的内容有十分深刻的认识。他是基于其切身的教学经验和研究成果，编写这本数据仓库领域的技术基础教材的。

本书涉及到了数据仓库技术中的各个领域，站在一个初学者的角度，论述了数据仓库的很多基本概念，对使用者的类型进行了划分，描述了建设数据仓库可能面临的各种风险，明确提出数据仓库承担着信息传递的功能的这一观点。书中具体论述了数据仓库技术的发展趋势，谈到具体实施过程的项目管理方法，还对数据仓库中的需求驱动给出了清晰的明确。此外，本书还对元数据管理给出了较详细的描述，提出了纬度建模的有关原则，同时论述了数据挖掘、OLAP 等技术，最后还描述了数据仓库项目是如何进行维护和升级的。

本书具有十分鲜明的特点，完全站在使用者的角度做深入浅出的技术论述，基本覆盖了数据仓库设计、建设、运营管理的各个方面，并提供了大量生动形象的图片，使读者非常容易理解。在技术论述上，有所侧重，突出了数据仓库的特点内容。

总体而言，本书是一本不可多得的介绍数据仓库基本技术的教材，对于国内数据仓库系统建设具有重要的指导作用，对学习数据仓库技术也将提供很好的帮助。

在本书的翻译过程中，感谢李剑威、韩洁、宋美娜做了大量的工作，翻译人员先后查阅了很多资料，对一些术语进行了仔细的推敲，段云峰博士进行了最后的定稿工作，在此对其他帮助过本书翻译过程的同仁表示衷心的感谢！

序

我很高兴能向信息技术的专业人士谈谈我对 Paulraj Ponniah 写的这本书——《数据仓库基础》的一些看法。1998 年春天, Raritan Valley 社区学院决定开一门关于数据仓库的课程。这源于长期教授数据库设计和开发课程的 Ponniah 博士的想法。但是当时很难找到一本在数据仓库知识方面非常适合学院课程的教材。我们不得不将就使用一本不很适合的书。为了教好这门课程, Ponniah 博士只好使用自己在数据仓库方面的研究资料对这本书进行了补充。我们的学生对这门课反映非常好, 而他们主要是当地信息技术专业人员。现在这本出色的数据仓库教科书终于问世了, 这要感谢 Ponniah 博士的卓识远见和辛勤工作, 以及富有洞察力的出版商——John Wiley&Sons 的支持。

这本书的独特优点有如下几点。

- 内容安排逻辑性强。
- 内容的取舍非常适合于作为介绍性的书籍, 而且涵盖面大。
- 主题经过了合理安排, 章节内容包括了数据仓库知识领域的所有重要方面, 从设计到应用过渡非常自然。
- 在每一章节中的主题连续性非常好。
- 本书中所包含的内容都是学习基础知识所必需的。
- 内容正确、紧跟潮流, 图表内容准确地强调和说明了主题。
- 每一章末尾都有丰富的复习思考题和练习。这在很多的数据仓库书中是找不到的。这些复习思考题和练习对于教师教学和学生学学习都非常有益。

Ponniah 博士的写作风格十分清晰简明。正是由于这本书的简洁和完备, 我相信它一定会得到大家的欢迎, 特别是高校的学生, 涉世未深的信息技术领域专业人员, 甚至是数据仓库专家。

虽然数据仓库方面的名家之作多如牛毛, 比如 Kimball, Inmon, Barquin 和 Singh 都有作品, 但是这本书有其独到之处, 信息技术的专业人员们一定会从此书中受益匪浅。此外, 本书应能得到大学教授及学习数据仓库课程的学生们欢迎。

总之, Ponniah 博士为学生和经验丰富的 IT 专业人员带来了这本好书。作为

一名长期从事信息技术教学的工作者，我向所有教授数据仓库课程的大学老师和
教研负责人，郑重地推荐此书。

PRATAP P.REDDY, Ph.D.

新泽西州拉里坦河学院北部支部

CIS 系系主任、教授

前 言

谁需要阅读本书

你是一直关注数据仓库领域发展变化的信息技术领域专业人员吗？你正在考虑涉足这一充满机会的新领域吗？你是渴望掌握数据仓库基础知识的系统分析员、程序员、数据分析师、数据库管理员、项目经理或者是软件工程师吗？你想知道你需要阅读多少本书才能学到数据仓库基础知识吗？你是否在数据仓库浩如烟海的文献和产品中迷失？你希望得到一本简单明了的，专门为信息技术领域专业人员量身定制的数据仓库知识的书吗？你需要一本深度恰到好处的教科书帮助你学习数据仓库基础知识吗？如果你对上面任何一个问题的回答是“是的”，那么这本书就很适合你。

这是一本专门为信息技术领域专业人员而写的权威书籍。本书的组织和内容陈述都是专门为信息技术领域专业人员而设计的，只对该领域感兴趣的业外人士不是本书的目标读者，编写此书旨在满足 IT 专业人士的具体需要。在内容上，没有特别强调某些特定的方面而忽略其他重要方面。这本书将把你带入一个完整的数据仓库的世界。

这本书是如何做到适应信息技术领域专业人员需要的呢？作为具有丰富 IT 行业实际经验的 IT 专业人员，作为一个成功的数据库和数据仓库的资深顾问，以及在大学课堂和公开研讨班里教授数据仓库基础课程的老师，我非常准确地了解 IT 专业人士的需要，在本书的每一章中，我都将谈到 IT 中的这些需求。

阅读背景

为什么这么多的公司都争相引入数据仓库技术？数据仓库已经不再是单纯的仅供科研和实验的概念，它已经成为一种主流。数据仓库固然还没普及到每个医生的办公室，但是它也并非仅限于高端商业的应用。超过半数以上的美国公司和很大一部分的全世界的公司都在使用数据仓库。

现在，从零售连锁店到金融机构，从生产企业到政府部门，从航空公司到公益事业，数据仓库正引起一场商业分析和战略决策方式的革命，每个拥有数据仓库的公司都正获取巨大利益，至少也得到了积极的结果。这些公司现在采用了基于 Web 的技术，增强了它们传递重要信息的潜力。

在过去的 5 年中，数以百计的供应商带着他们的数据仓库产品涌入这个市场，他们提供的解决方案和产品已经包含了数据仓库的整个产品线——数据建模、数据获取、数据质量、数据分析、元数据等。这一市场已经很大，并且还在继续增长。

IT 角色的改变

在这样的情况下，在所有成长型公司中，信息技术部门感受到了自己角色的根本性变化。IT 不再需要创造每一份报告，给最终用户提供信息。现在的 IT 部门的职责是建造信息传递系统，使最终用户自己通过新的方法来得到分析和决策用的信息。数据仓库被证明是一种成功的信息传递系统。

负责建造数据仓库的 IT 专业人员需要转变他们关于开发应用程序的观念。他们得明白，数据仓库的某种设计并非放诸四海皆准，他们必须清楚了解如何从源数据系统里抽取、转换、展现数据，以及数据仓库的体系、结构，甚至各种信息传递的方法。

简言之，像您这样的 IT 专业人士必须牢固掌握数据仓库基础知识。

本书可以为你做什么

本书全面又详细，你可以学到有关数据仓库每一个重要部分的内容，包括计划、需求、体系、基础、结构、设计、数据准备、信息传递、配置和维护等。本书专为 IT 专业人员设计：你可以很容易跟上进度，因为它是根据 IT 专业人员的背景、知识而设计的，而你对那些技术术语又非常熟悉。本书编排合理，从对概念的总体介绍入手，然后介绍了计划和需求，之后是结构和基本体系，再到数据设计，信息传递，最后是数据仓库的配置和维护。这一推进过程是你的经验和每

天的工作中所熟悉的。

本书还提供了互动的学习方式。它不是单纯的讲授，你可以通过每一章末的复习思考题和练习来参与学习。在每一章中，目标，提出本章主题，章后小结列出本章包含的内容，你可以将每一个概念和技术同数据仓库的实践和市场结合起来。你将会发现很多实际应用中的例子。虽然本书是你踏入数据仓库领域里的第一本基础书籍，但是它几乎涵盖该领域所有方面的知识，因此将使你在数据仓库领域向成为专家的目标迈出非常坚实的一步。

另外，本书还非常适合作为教科书，供自学人员、大学课程或者科研机构使用。它将为你提供成为数据仓库领域专家的机会。

在这里，我要感谢我在本书末尾提到的作者们，他们的洞察力和观察力帮助我为本书选择到合适且足够的内容。我也必须向我的学生和同事们表示感谢，大家的合作帮助我使这本书更好地适合 IT 专业人员的需要。

Paulraj Ponniah, Ph.D.

于新泽西州

目 录

第 1 章 对数据仓库的迫切需求	1
本章目标	1
1.1 对战略信息的不断增长的需求	2
1.1.1 信息危机	4
1.1.2 技术趋势	5
1.1.3 机遇和风险	6
1.2 以往的决策支持系统的失败	8
1.2.1 决策支持系统的历史	9
1.2.2 缺乏提供战略信息的能力	10
1.3 操作型系统和决策支持系统	11
1.3.1 使商业运作起来	11
1.3.2 监视商业的运作	12
1.3.3 不同的范围,不同的目的	12
1.4 数据仓库——惟一可行的解决方案	13
1.4.1 一种新类型的系统环境	13
1.4.2 新环境下的需求处理	14
1.4.3 数据仓库的商业智能	14
1.5 数据仓库的定义	15
1.5.1 数据仓库是信息传递的一种简单概念	15
1.5.2 数据仓库是一种环境,而不是产品	16
1.5.3 数据仓库是多种技术的混合体	16
本章小结	17
复习思考题	18
练习题	18
第 2 章 数据仓库的组成部分	20
本章目标	20
2.1 定义的特点	21
2.1.1 面向主题的数据	21
2.1.2 综合的数据	22
2.1.3 数据的时间特性	24
2.1.4 数据的非易变性	24
2.1.5 数据粒度	26
2.2 数据仓库和数据集市	27
2.2.1 它们有什么不同	27

2.2.2	自上而下和自下而上的方法	28
2.2.3	一个实用的方法	29
2.3	数据仓库的组成部分	30
2.3.1	源数据部分	31
2.3.2	数据准备部分	34
2.3.3	数据存储部分	37
2.3.4	信息传递部分	37
2.3.5	元数据部分	38
2.3.6	管理和控制部分	39
2.4	数据仓库中的元数据	39
2.4.1	元数据的类型	39
2.4.2	元数据的特殊意义	40
	本章小结	41
	复习思考题	41
	练习题	42
第3章	数据仓库的发展趋势	43
	本章目标	43
3.1	数据仓库的持续成长	44
3.1.1	数据仓库正在成为主流	44
3.1.2	数据仓库的扩张	45
3.1.3	解决方案和产品	46
3.2	重要趋势	48
3.2.1	多种数据类型	48
3.2.2	数据可视化	51
3.2.3	并行处理	53
3.2.4	查询工具	55
3.2.5	浏览工具	55
3.2.6	数据融合	56
3.2.7	多维分析	56
3.2.8	代理技术	57
3.2.9	企业组合数据	57
3.2.10	数据仓库和 ERP	58
3.2.11	数据仓库和知识管理	59
3.2.12	数据仓库和 CRM	60
3.2.13	活跃的数据仓库	62
3.3	标准的出现	62
3.3.1	元数据	63
3.3.2	OLAP	64
3.4	支持 Web 的数据仓库	65
3.4.1	将数据仓库放入 Web 中	65
3.4.2	将 Web 技术引入数据仓库	66

3.4.3 支持 Web 技术的配置	67
本章小结	68
复习思考题	68
练习题	69
第 4 章 规划和项目管理	70
本章目标	70
4.1 规划你的数据仓库	71
4.1.1 关键问题	71
4.1.2 商业需求, 而非技术	74
4.1.3 高层管理的支持	75
4.1.4 数据仓库的可行性分析	75
4.1.5 全盘计划	77
4.2 数据仓库项目	77
4.2.1 有什么不同	78
4.2.2 准备情况的评估	79
4.2.3 生命周期方法	80
4.2.4 开发的各阶段	82
4.3 项目团队	83
4.3.1 组织项目团队	83
4.3.2 角色和责任	84
4.3.3 技能和经验水平	87
4.3.4 用户参与	88
4.4 项目管理要素	89
4.4.1 项目管理的指导原则	90
4.4.2 警告征兆	91
4.4.3 成功的因素	92
4.4.4 成功项目细审	93
4.4.5 采用实用的方法	94
本章小结	95
复习思考题	96
练习题	96
第 5 章 定义商业需求	98
本章目标	98
5.1 维度分析	98
5.1.1 对不可预知的信息的使用	99
5.1.2 商业数据的维度	99
5.1.3 商业维度的例子	101
5.2 信息包——一个新概念	102
5.2.1 不完全确定的需求	102
5.2.2 商业维度	104
5.2.3 维度层次和分类	104

5.2.5 关键商业指标或事实	106
5.3 收集需求的方法	108
5.3.1 采访技巧	109
5.3.2 调整联合应用程序设计方法	111
5.3.3 回顾已有的文档	113
5.4 需求定义: 范围和内容	114
5.4.1 数据源	115
5.4.2 数据转换	115
5.4.3 数据存储	116
5.4.4 信息传递	116
5.4.5 信息包表	116
5.4.6 需求定义文档提纲	116
本章小结	117
复习思考题	117
练习题	118
第6章 需求——数据仓库的驱动力	120
本章目标	120
6.1 数据设计	121
6.1.1 商业维度的结构	123
6.1.2 关键衡量指标的结构	124
6.1.3 细节层次	124
6.2 体系结构规划	125
6.2.1 组成部分的构成	126
6.2.2 特殊因素	127
6.2.3 工具和产品	130
6.3 数据存储规范	132
6.3.1 数据库管理系统的选择	133
6.3.2 存储规模估计	134
6.4 信息传递策略	135
6.4.1 查询和报表	136
6.4.2 分析的类型	137
6.4.3 信息分发	137
6.4.4 决策支持应用程序	137
6.4.5 发展和扩大	138
本章小结	138
复习思考题	138
练习题	139
第7章 体系结构及其组成部分	141
本章目标	141
7.1 掌握数据仓库的体系结构	141
7.1.1 体系结构: 定义	141

7.1.2 三个主要区域的体系结构	142
7.2 区别于其他结构的特点	143
7.2.1 不同的目标和范围	144
7.2.2 数据内容	145
7.2.3 复杂分析和快速响应	145
7.2.4 灵活性和动态性	146
7.2.5 元数据驱动	146
7.3 体系结构框架	147
7.3.1 支持数据流的体系结构	147
7.3.2 管理和控制模块	148
7.4 技术体系结构	149
7.4.1 数据获取	151
7.4.2 数据存储	154
7.4.3 信息传递	156
本章小结	158
复习思考题	159
练习题	159
第8章 数据仓库的基础构造	161
本章目标	161
8.1 支持体系结构的基础构造	161
8.1.1 操作型基础构造	163
8.1.2 物理基础构造	163
8.2 硬件和操作系统	165
8.2.1 平台选择	166
8.2.2 服务器硬件	178
8.3 数据库软件	183
8.3.1 并行处理方案	184
8.3.2 数据库管理系统的选择	186
8.4 工具收集	187
8.4.1 先设计好体系结构,再选择工具	188
8.4.2 数据建模	189
8.4.3 数据抽取	189
8.4.4 数据转换	189
8.4.5 数据装载	190
8.4.6 数据质量	190
8.4.7 查询和报表	190
8.4.8 联机分析处理(OLAP)	190
8.4.9 预警系统	191
8.4.10 中间件及连接部件	191
8.4.11 数据仓库管理	191
本章小结	191

复习思考题	192
练习题	192
第9章 元数据的重要角色	194
本章目标	194
9.1 元数据的重要性	194
9.1.1 数据仓库的关键需求	196
9.1.2 为什么元数据对最终用户很关键	200
9.1.3 为什么元数据对 IT 人员很关键	201
9.1.4 数据仓库任务自动化	203
9.1.5 建立信息上下文	205
9.2 按功能区域划分的元数据类型	206
9.2.1 数据获取	207
9.2.2 数据存储	208
9.2.3 信息传递	210
9.3 商业元数据	211
9.3.1 内容总揽	211
9.3.2 商业元数据举例	212
9.3.3 内容重点	213
9.3.4 谁会受益	213
9.4 技术元数据	214
9.4.1 内容总揽	214
9.4.2 技术元数据举例	214
9.4.3 内容重点	216
9.4.4 谁会受益	216
9.5 如何提供元数据	217
9.5.1 元数据需求	218
9.5.2 元数据的来源	219
9.5.3 元数据管理面临的挑战	221
9.5.4 元数据储存库	222
9.5.5 元数据集成与标准	224
9.5.6 实施选项	224
本章小结	226
复习思考题	227
练习题	227
第10章 维度建模的原则	229
本章目标	229
10.1 从需求到数据设计	229
10.1.1 设计决策	230
10.1.2 维度建模基础	230
10.1.3 E-R 建模与维度建模的对比	235
10.1.4 使用 CASE 工具	236

10.2 星型模式	237
10.2.1 一个简单的星型模式的回顾	237
10.2.2 维度表的内容	240
10.2.3 事实表的内容	242
10.2.4 不含事实的事实表	244
10.2.5 数据粒度	245
10.3 星型模式的键	246
10.3.1 主键	246
10.3.2 替代键	247
10.3.3 外键	248
10.4 星型模式的优势	248
10.4.1 用户容易理解	249
10.4.2 优化浏览	250
10.4.3 最适于查询处理	251
10.4.4 星型连接和星型索引	252
本章小结	252
复习思考题	253
练习题	253
第 11 章 维度建模：高级专题	255
本章目标	255
11.1 维度表的更新	256
11.1.1 慢速变化中的维度	256
11.1.2 第 1 类修改：改正错误	257
11.1.3 第 2 类修改：保存历史数据	259
11.1.4 第 3 类修改：暂时的（软性的）修改	260
11.2 各式各样的维度	262
11.2.1 大维度	263
11.2.2 快速变化中的维度	264
11.2.3 废弃维度	266
11.3 雪花型模式	267
11.3.1 规范化选项	268
11.3.2 优势与劣势	269
11.3.3 什么时候使用雪花型模式	270
11.4 聚集事实表	271
11.4.1 事实表的大小	273
11.4.2 对聚集的需求	275
11.4.3 对事实表进行聚集	275
11.4.4 聚集的选项	281
11.5 星型模式族	283
11.5.1 快照表和事务表	284
11.5.2 核心表和定制表	285

11.5.3 支持企业价值链或者价值环	286
11.5.4 使维度一致	287
11.5.5 将事实表标准化	288
11.5.6 星型模式族小结	289
本章小结	290
复习思考题	290
练习题	291
第 12 章 数据抽取、转换和装载	292
本章目标	292
12.1 ETL 概览	293
12.1.1 最重要和最具有挑战性	294
12.1.2 耗时而且费劲	295
12.1.3 ETL 的需求和步骤	296
12.1.4 关键因素	297
12.2 数据抽取	298
12.2.1 数据源确认	299
12.2.2 数据抽取技术	300
12.2.3 技术的评估	307
12.3 数据转换	309
12.3.1 数据转换：基本任务	310
12.3.2 主要转换类型	311
12.3.4 数据整合和合并	313
12.3.5 维度属性的转换	315
12.3.6 如何实施转换	316
12.4 数据装载	318
12.4.1 应用数据：技术和过程	319
12.4.2 数据的刷新和更新	322
12.4.3 维度表的规程	323
12.4.4 事实表：历史与增量的装载	324
12.5 ETL 总结	325
12.5.1 ETL 工具选项	326
12.5.2 再次强调 ETL 中的元数据 (Metadata)	327
12.5.3 ETL 的总结和方法	328
本章小结	329
复习思考题	330
练习题	330
第 13 章 数据质量：成功的关键	332
本章目标	332
13.1 为什么数据质量如此重要	333
13.1.1 什么是数据质量	334
13.1.2 提高数据质量的好处	337

13.1.3 数据质量问题的类型	338
13.2 数据质量的挑战	341
13.2.1 数据污染的来源	342
13.2.2 姓名和地址的有效性	344
13.2.3 数据质量低劣带来的代价	345
13.3 数据质量工具	346
13.3.1 数据清洗工具的分类	346
13.3.2 错误发现特性	346
13.3.3 数据修正特性	347
13.3.4 数据库管理系统的质量控制	347
13.4 确保数据质量的第一步	348
13.4.1 数据清洗的决策	349
13.4.2 谁应该负责	352
13.4.3 净化过程	353
13.4.4 对数据质量的实用建议	355
本章小结	356
复习思考题	356
练习题	357
第 14 章 信息和用户类型之间的匹配	359
本章目标	359
14.1 来自于数据仓库的信息	360
14.1.1 数据仓库和操作型系统	360
14.1.2 信息潜力	362
14.1.3 用户—信息接口	366
14.1.4 行业应用	368
14.2 谁将使用这些信息	368
14.2.1 用户的种类	369
14.2.2 他们需要什么	372
14.2.3 怎样为用户提供信息	376
14.3 信息传递机制	377
14.3.1 查询	379
14.3.2 报表	381
14.3.3 分析	382
14.3.4 应用程序	383
14.5 信息传递工具	384
14.5.1 桌面电脑环境	384
14.5.2 工具选择的方法学	385
14.5.3 选择工具的标准	389
14.5.4 信息传递框架	390
本章小结	391
复习思考题	392