

HZ BOOKS
华章科技

Mc
Graw
Hill Education

商务智能与信息化技术丛书

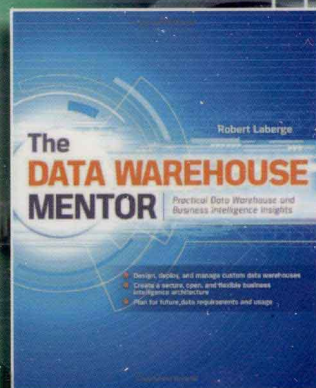
设计、部署和管理自定义数据仓库
创建安全、开放和灵活的商务智能架构
规划未来数据需求和使用

数据仓库应用指南

数据仓库与商务智能最佳实践

The Data Warehouse Mentor

Practical Data Warehouse and Business Intelligence Insights



Robert Laberge 著
祝洪凯 李妹芳 译



机械工业出版社
China Machine Press

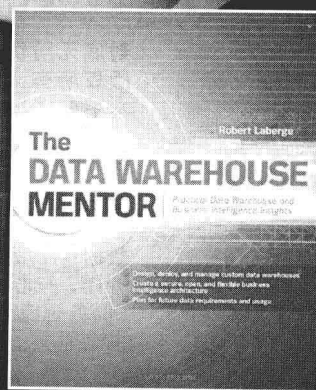
商务智能与信息化技术丛书

数据仓库应用指南

数据仓库与商务智能最佳实践

The Data Warehouse Mentor

Practical Data Warehouse and Business Intelligence Insights



Robert Laberge 著
祝洪凯 李妹芳 译



机械工业出版社
China Machine Press

Robert Laberge: The Data Warehouse Mentor; Practical Data Warehouse and Business Intelligence Insights (ISBN: 978-0-07-174532-1).

Copyright © 2011 by The McGraw-Hill Companies, Inc.

All Rights reserved. No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including without limitation photocopying, recording, taping, or any database, information or retrieval system, without the prior written permission of the publisher.

This authorized Chinese translation edition is jointly published by McGraw-Hill Education (Asia) and China Machine Press. This edition is authorized for sale in the People's Republic of China only, excluding Hong Kong, Macao SAR and Taiwan.

Copyright © 2012 by McGraw-Hill Education (Asia), a division of the Singapore Branch of The McGraw-Hill Companies, Inc. and China Machine Press.

版权所有。未经出版人事先书面许可，对本出版物的任何部分不得以任何方式或途径复制或传播，包括但不限于复印、录制、录音，或通过任何数据库、信息或可检索的系统。

本授权中文简体字翻译版由麦格劳-希尔（亚洲）教育出版公司和机械工业出版社合作出版。此版本经授权仅限在中华人民共和国境内（不包括香港特别行政区、澳门特别行政区和台湾）销售。

版权© 2012 由麦格劳-希尔（亚洲）教育出版公司与机械工业出版社所有。

本书封面贴有 McGraw-Hill 公司防伪标签，无标签者不得销售。

封底无防伪标均为盗版

版权所有，侵权必究

本书法律顾问 北京市展达律师事务所

本书版权登记号：图字：01-2011-3603

图书在版编目（CIP）数据

数据仓库应用指南：数据仓库与商务智能最佳实践/拉伯格（Laberge, R.）著；祝洪凯，李妹芳译. —北京：机械工业出版社，2012.3

（商务智能与信息化技术丛书）

书名原文：The Data Warehouse Mentor; Practical Data Warehouse and Business Intelligence Insights

ISBN 978-7-111-37044-4

I. 数… II. ①拉… ②祝… ③李… III. 数据库系统 IV. TP311.13

中国版本图书馆 CIP 数据核字（2012）第 002411 号

机械工业出版社（北京市西城区百万庄大街 22 号 邮政编码 100037）

责任编辑：吴怡

北京京师印务有限公司印刷

2012 年 3 月第 1 版第 1 次印刷

186mm × 240mm · 21 印张

标准书号：ISBN 978-7-111-37044-4

定价：69.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换

客服热线：(010) 88378991；88361066

购书热线：(010) 68326294；88379649；68995259

投稿热线：(010) 88379604

读者信箱：hzsj@hzbook.com

译者序

数据仓库和商务智能在各个领域的应用已经如火如荼。本书作者给我们分享了他 30 多年的工作经验。本书涉及面广泛，内容全面，从基础概念的介绍、各个组件的剖析，到实践中的问题，作者都给出了细致的描述，深入浅出、高屋建瓴地阐述了数据仓库和商务智能的方方面面。本书理论和实践相结合，是一本不错的数据仓库和商务智能方面的整体指南。

本书分为三个部分来讲解数据仓库这个复杂系统，以及实现商务智能的有效方案。第一部分介绍了关于商务智能和数据仓库的基础概念，旨在介绍基础知识，为管理者思考为何、如何建立数据仓库提供了思考的方向。第二部分介绍数据仓库系统的基础组件，这部分涉及数据仓库和商务智能系统的技术方面，探讨了如何建立数据仓库系统，来维护企业资产并提供商务智能支持工作。第三部分从实践角度说明了如何构建数据仓库系统，包括经典的构建场景以及后期工作。

说实话，业余翻译了几本书，总是很担心译得不好误导了读者。在翻译的过程中我们查阅了大量文献和网络资源，为了译好一些不常见的词汇也反复琢磨了作者原意。但是还是觉得时间紧迫，翻译仓促，惶恐不安。在这里要感谢机械工业出版社华章分社编辑吴怡老师的很多辛苦付出，也感谢所有其他为本书付出努力的编辑们。

由于时间、精力、能力有限，本书的疏漏、错误之处在所难免，还望各位读者不吝指正。

前 言

本书对数据仓库世界中很多主题进行了探讨。本书旨在从业务和技术角度说明数据仓库系统的构建，侧重于简单朴实地描述如何构建切实的解决方案。这些见解来源于我30多年在20多个国家中50多家企业的亲身经历，在这些经历中，我曾作为独立顾问、员工以及IBM产业模式和资产实验室的合伙人，见证了很多数据仓库的实施过程。

本书介绍了构建数据仓库的组件和不同选择，以及选择某种方式的利弊。每家企业的数据仓库构建都是具有其特色的，但可以借鉴全球范围内很多企业的各种数据仓库和商务智能环境中获取的知识。本书首先从高层角度介绍了数据仓库主题，以确保对术语和上下文理解一致，然后详细说明了各个主题。这些主题都和数据仓库、商务智能和性能管理相关。

对于数据仓库的构建不存在规则，但是有很多指南。本书的主要根本点是根据具体的和对业务需求的理解，构建适应特定企业需求的解决方案，同时为今后的工作创建一个开放、灵活的架构基础。很多企业在初始包含商务智能报表的集中式数据仓库的构建上花费了大量的预算，结果却发现其创建的解决方案过于具体，只适合一两个用途，而无法满足后期的需求。当然，我们无法对未来进行预测，但是可以在一定程度上预期今后的数据需求和使用方式，确保设计和构建环境灵活、开放，对于变化可扩展而不需要每次重新设计和构建。

很多企业的领导人意识到企业数据是企业的基础资产，必须对它进行组织、结构化和维护，以保证其业务信息有较好的质量和管理，从而在整个企业范围内共享。如果没有信息系统，企业就无法运作，而如果没有商业目的，信息系统就不复存在。它们相互依存，应该充分意识到信息架构和使用方式，以使得企业变得更加智慧。

本书结构

第一部分：准备

第一部分介绍商务智能和数据仓库的基础概念，旨在介绍基础知识，为管理工

作奠定基础。

第1章：数据仓库和商务智能概述 该章概要介绍了商务智能和数据仓库，最后提出了和数据仓库实现相关的高层次问题。

第2章：企业中的数据 该章探讨了数据如何作为企业资产，并提出关于如何组织数据的见解。

第3章：为什么创建数据仓库 该章探讨了支持和反对构建数据仓库的各种理由。“支持”的理由在于已经有一些构建数据仓库的经典场景，而“反对”的理由在于企业的文化和局限性能否推动项目向前发展。

第4章：数据仓库和商务智能战略 该章给出了构建数据仓库和商务智能行动的一些规划，探讨从何处以及如何启动项目，这取决于这项工作是面向商业报表解决方案，还是努力将数据进行组织和结构化。

第5章：项目资源：角色和洞察力 该章讨论了数据仓库项目的关键角色，以及最佳实践的团队结构。

第6章：项目总结概论 该章简要介绍了项目章程、项目范畴和工作说明书的内容。

第二部分：组件

第二部分介绍了数据仓库系统的基础组件，深入分析了数据仓库和商务智能系统的技术方面。这部分具体探讨了数据仓库系统中用以维护企业资产和提供商务智能支持工作的各个组件。

第7章：商务智能：数据集市及其使用方式 该章从数据模型到性能问题，详细探讨了数据集市及其使用方式。

第8章：企业数据模型 该章讨论了企业数据模型、如何构建企业数据模型的一些实例以及一般问题。

第9章：数据仓库架构：组件 该章从建模和数据流角度探讨了数据仓库架构的不同类型。

第10章：ETL和数据质量 该章探讨了数据仓库中的数据采集层和分发层的一些普遍特征，并提出关于数据质量问题的一些见解。

第11章：项目规划和方法论 该章讨论了数据仓库和商务智能项目规划的一些方法。

第三部分：构建

第三部分从实践角度说明了如何构建数据仓库系统。这部分旨在介绍经典的构建场景和工作，以及数据监理和对后期工作的审查。

第 12 章：工作场景 该章介绍了如何使用自上而下、自下而上和混合式方法来构建数据仓库和商务智能系统，并讨论了一些其他主题，包括简要介绍企业信息架构。

第 13 章：数据监理 该章探讨了企业数据监理，包括企业结构、数据质量、所有权和变更管理。

第 14 章：项目后评审 该章探讨了数据仓库和商务智能项目在开发完成后的一些方面。

本书力争做到成为构建数据仓库系统的完整指南，目标是理解当今数据仓库系统中的很多问题，并从多个角度提出自己的观点。作者希望本书能够帮助你构建好自己的数据仓库。

希望你喜欢本书！

致谢

特别感谢本书的技术编辑 David Marcotte 和 Ken Yu，感谢他们帮我审查书稿，并提出他们的观点和建议。他们的宝贵意见为本书最终的成功出版提供了方向指引。真心感激！

David Marcotte 是全球性零售业和分析业的大师，他拥有该行业的渊博知识。

Ken Yu 是数据设计和数据流方面的技术专家，他在很多部门工作过，积累了对数据建模和数据架构的很多实践经验和常识性方法。

特别感谢我的妻子 Rakhee Laberge（工商管理硕士），感谢她对本书的审查以及提出的很多意见和建议。

Robert (Bob) Laberge

联系方式：datawarehousementor@gmail.com

作者简介

Robert (Bob) Laberge 是多家互联网企业的创始人、IBM 产业模式和资产实验室的首席顾问，他的研究重点是数据仓库和商务智能解决方案。

Bob 早在 20 世纪 70 年代末就开始其职业生涯，当时比尔·盖茨还只是一个百万富翁，Bob 曾经是开发人员、数据仓库管理员、数据建模师、项目经理、数据架构师、企业信息架构师、数据仓库/商务智能审计员、战略师，而且还是富于创新的企业家。从那时，Bob 就跑遍全球，通过设计、优化、最佳实践和在概念层、逻辑层和物理层的常识说明，提供指导、培训和证明数据仓库和商务智能实践经验和解决方案。Bob 成功地帮助了 50 多家大型企业扩展业务，这些企业涉及零售、保险、医疗、铁路、电信、电子商务和银行等行业。

Bob 拥有英国 Durham 大学的工商管理硕士学位。你可以通过 datawarehousementor@gmail.com 联系他。



技术编辑简介

David Marcotte 是 Kantar Retail 的美洲零售分析的高级副总裁，其专业领域涉及国际销售、采购、零售业、供应链和商务智能。他在很多技术主题上发表过演讲，并且是高深层会议的培训师，和全球范围的众多财富 500 强公司合作。最近，他是可口可乐零售研究会的《2010 年零售业新兴市场研究》的研究人员和作者，该研究总结了在中国、秘鲁、土耳其、波兰、南非和巴西的调查结果。在加入 Kantar Retail 之前，他是 IBM 的全球商务智能组的负责人。在 IBM 工作期间，他参与开发团队，提供零售业的一系列解决方案，包括损失预防、定价、分类管理过程、商品规划、供应链、商品销售和销售团队优化。他还是 IBM 零售数据仓库解决方案的业务和项目负责人，在这个项目中，他首次和本书的作者 Bob Laberge 合作。他为很多

的零售商和制造商提供解决方案，这些公司包括索尼、Sobeys、D&S、Falabella、Shoppers Drug Mart 和 Kroger 等。David 在很多行业委员会和专题讨论组工作过，包括 IRI 产品咨询委员会和高效消费者响应组（ECR）。你可以通过 David. Marcotte@KantarRetail.com 和他联系。

Ken Yu 在信息管理和数据架构领域拥有超过 25 年的工作经验。他曾经在不同行业和应用中担任过企业信息架构师、项目架构师、数据建模师、数据仓库管理员和开发人员。1995 年，Ken 加入 IBM 以支持大型数据库客户端工作，担任业务负责人。作为信息管理顾问，他在整个企业范围内提供关于数据监理、主数据和元数据管理、数据建模和参考模型的价值等主题的指导和培训。他一直是众多关键任务应用以及商务智能和数据仓库解决方案的主要设计、开发、技术架构和业务过程参与人员。当前，Ken 是 Cypselurus 公司的首席顾问，该公司在加拿大安大略省的多伦多市。

目 录

译者序
前言
作者简介

第一部分 准 备

第 1 章 数据仓库和商务智能概述	2
1.1 商务智能概述	2
1.1.1 定义	3
1.1.2 商务智能的价值	4
1.1.3 剖析商务智能	5
1.1.4 商务智能的成功要素	6
1.1.5 商务智能的目标	7
1.1.6 BI 用户展现层	9
1.1.7 BI 工具和架构	12
1.1.8 全球化带来的发展	14
1.2 数据仓库概述	14
1.2.1 定义	14
1.2.2 数据仓库系统	15
1.2.3 数据仓库架构	16
1.2.4 数据流术语	18
1.2.5 数据仓库目标	20
1.2.6 数据结构化策略	22
1.2.7 数据仓库业务	23
1.3 常见问题	24
1.3.1 当前系统是否足够好	24

1.3.2	数据仓库的价值	25
1.3.3	成本多高	26
1.3.4	时间多长	27
1.3.5	成功的因素	29
第2章	企业中的数据	33
2.1	企业资产	33
2.1.1	具有上下文的数据	33
2.1.2	数据质量	35
2.1.3	数据字典	37
2.1.4	数据组件	38
2.2	组织数据	42
2.2.1	对数据结构化	43
2.2.2	数据模型	44
2.2.3	数据架构	48
2.3	竞争优势	52
2.3.1	构建还是购买数据模型	53
2.3.2	指导业务	56
第3章	为什么创建数据仓库	58
3.1	平台迁移	59
3.1.1	业务连续性	60
3.1.2	逆向工程	60
3.1.3	数据质量	61
3.1.4	并行环境	62
3.1.5	附加值	63
3.2	数据仓库集中化	63
3.2.1	企业间并购	63
3.2.2	企业内合并	64
3.2.3	集中式设计和局部使用	64
3.3	数据集市整合	64
3.4	新方案	66

3.5	新方案：动态报表	68
3.6	“Just Build It” 模式	69
3.7	数据 Floundation	71
3.8	不构建数据仓库的原因	72
3.8.1	数据质量差	72
3.8.2	缺乏商业目标	73
3.8.3	缺乏管理层支持	73
3.8.4	目标不明确	73
3.8.5	当前系统足够用	73
3.8.6	缺乏人才资源	74
3.8.7	环境不稳定	74
3.8.8	成本太高	74
3.8.9	管理不善	74
第4章	数据仓库和商务智能战略	75
4.1	商务智能战略	75
4.1.1	商业目标	75
4.1.2	商业用途	76
4.1.3	架构概览	77
4.2	数据仓库战略	79
4.2.1	用途	79
4.2.2	数据仓库架构	80
4.3	重点和成功	83
4.3.1	整个企业还是业务线	83
4.3.2	目标明确	84
4.3.3	成功：衡量的标准是什么	84
4.4	从何处着手	85
4.4.1	关于商务智能	86
4.4.2	关于数据仓库	87
4.5	如何开始	87
4.5.1	关于商务智能	87
4.5.2	关于数据仓库	90

4.6	项目阶段化	92
4.7	需要多长时间（重新回顾）	93
4.8	兴趣点	95
4.8.1	常见的失败原因	95
4.8.2	基本原则	99
第5章	项目资源：角色和洞察力	101
5.1	关键点	101
5.1.1	项目团队	102
5.1.2	资深专业知识	102
5.1.3	领导力	103
5.1.4	项目发起人	105
5.1.5	数据仓库管理层	105
5.2	团队结构	106
5.2.1	管理层发起人	107
5.2.2	数据管家	108
5.2.3	基本资源	108
5.3	定期审查：进度审核	112
5.4	能力中心	112
第6章	项目总结概论	114
6.1	项目章程	114
6.2	项目范畴	116
6.3	工作说明书	117

第二部分 组 件

第7章	商务智能：数据集市及其使用方式	120
7.1	为什么要对数据建模	121
7.1.1	数据模型的类型	122
7.1.2	数据设计	125
7.2	事实表	132

7.2.1	事实的类型	133
7.2.2	事实表的类型	135
7.2.3	衡量指标来源	137
7.2.4	事实表关键字	137
7.2.5	事实表粒度	138
7.2.6	事实表密度	138
7.2.7	无事实的事实表	138
7.3	维度表	139
7.3.1	维度还是指标	140
7.3.2	历史表和日期表	141
7.3.3	维度表关键字	143
7.3.4	维度表的粒度	145
7.3.5	维度属性的来源和价值	146
7.3.6	维度类型	148
7.3.7	级别和辅助表	156
7.3.8	个人信息表	158
7.3.9	维度数	160
7.4	规模	160
第8章	企业数据模型	162
8.1	数据模型概览	162
8.2	构建企业数据模型的目标	166
8.3	企业数据模型的好处	166
8.4	数据模型：从何处开始	167
8.5	完全自上而下的数据模型	167
8.5.1	主题领域模型	168
8.5.2	概念模型	171
8.5.3	实体关系模型	171
8.6	总线结构	172
8.7	购买的数据模型	174
8.8	模型分析	176
8.8.1	数据组件	176

8.8.2	范化数据模型	178
8.8.3	超类和子类模型	182
8.8.4	在范化的数据模型中收集历史信息	185
8.8.5	代理键	189
8.8.6	逻辑和物理数据模型	190
8.8.7	是否具备参照完整性	191
8.9	其他数据模型	192
8.9.1	输入数据模型	192
8.9.2	临时存储数据模型	192
8.10	最后的思考	193
第9章	数据仓库架构：组件	194
9.1	架构概述	194
9.2	架构师角色	195
9.2.1	解决方案架构师	195
9.2.2	数据仓库架构师	195
9.2.3	技术架构师	196
9.2.4	数据架构师	196
9.2.5	ETL架构师	196
9.2.6	BI架构师	196
9.2.7	综合	197
9.3	体系结构分层	198
9.3.1	单层体系结构	198
9.3.2	经典的两层体系结构	199
9.3.3	高级的三层体系结构	200
9.4	数据仓库架构	201
9.4.1	单独的数据集市架构	201
9.4.2	总线结构	202
9.4.3	中央存储库架构	202
9.4.4	联合架构	203
9.5	组件（分层）	204
9.5.1	数据源	204

9.5.2	数据生成	205
9.5.3	数据组织	205
9.5.4	数据分发	205
9.5.5	信息输出	205
9.6	实现方式	206
9.6.1	数据设计和数据流	207
9.6.2	逻辑和物理模型	207
9.6.3	自上而下的方式	209
9.6.4	自下而上的方式	211
9.6.5	混合模式	212
9.7	捷径	213
9.7.1	数据采集层	214
9.7.2	中央数据层	214
9.7.3	数据分发层	215
9.7.4	表现层	215
9.7.5	用户展现层	215
9.7.6	方法论	216
9.7.7	现成的解决方案	216
第 10 章	ETL 和数据质量	217
10.1	架构	218
10.1.1	数据获取	219
10.1.2	数据分发	220
10.1.3	ETL 映射	221
10.1.4	初始加载和增量加载	223
10.1.5	ETL、ELT 和 ETTL	224
10.1.6	并行操作	225
10.1.7	ETL 功能角色	226
10.1.8	数据流图	227
10.1.9	业务数据存储系统	228
10.2	数据源系统	229
10.2.1	没有数据源	229

10.2.2	多个数据源	229
10.2.3	其他来源（结构化输入文件）	230
10.2.4	非结构化数据	230
10.3	数据剖析	232
10.4	数据获取	232
10.4.1	多个大文件	233
10.4.2	伪文件	233
10.4.3	故障预防策略	234
10.5	转换和临时数据存储	234
10.5.1	准备工作	235
10.5.2	代理键	237
10.5.3	参照完整性	239
10.5.4	聚合、分析和汇总	240
10.5.5	编码表	240
10.6	加载	240
10.6.1	是否加载历史数据	241
10.6.2	插入、更新、插入或更新、删除	241
10.6.3	数据获取信息	242
10.6.4	加载调度	242
10.7	企业数据仓库的临时数据存储和总线架构的临时数据存储	242
10.8	数据分发	244
10.9	数据质量	246
10.10	ETL 工具	247
第 11 章	项目规划和方法论	248
11.1	基础	250
11.1.1	风险：逐步发展	251
11.1.2	风险：数据质量	251
11.1.3	风险：资源	252
11.1.4	风险：成本	253
11.1.5	变更管理	253
11.1.6	最佳实践	254