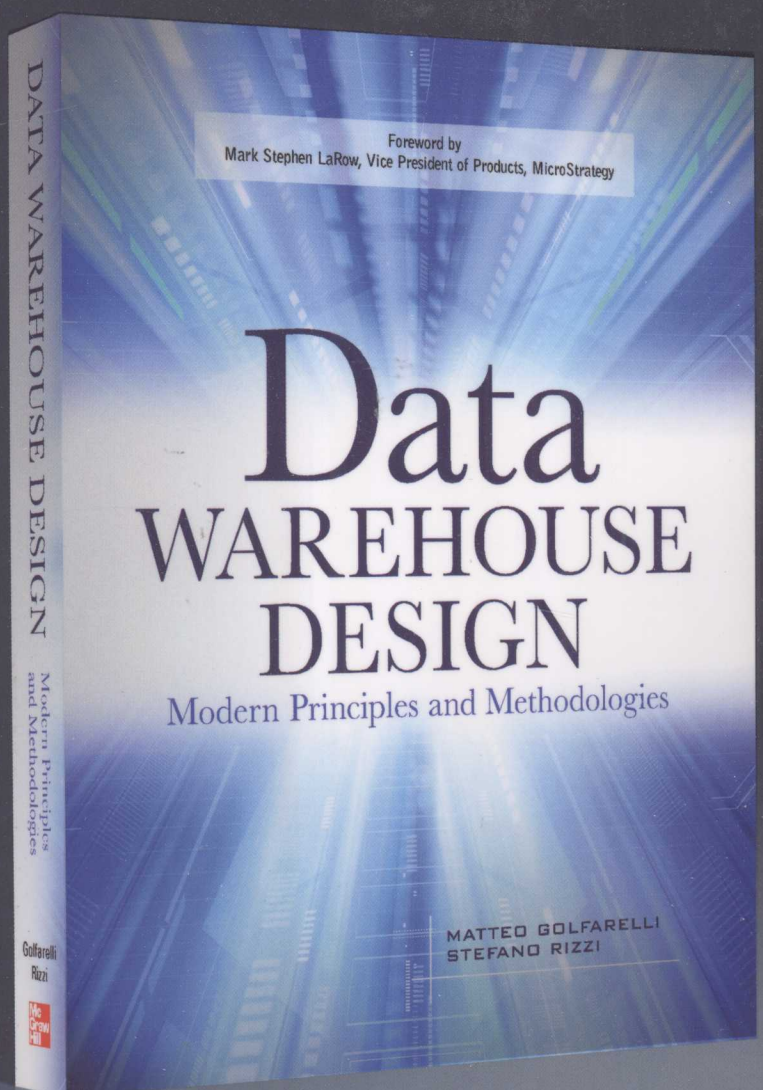
国外计算机科学经典教材

数据仓库设计：现代原理与方法

(意大利) Matteo Golfarelli
Stefano Rizzi 著

战晓苏 吴云浩 皮人杰 译



Data Warehouse Design:
Modern Principles and Methodologies

McGraw Hill Education

清华大学出版社

国外计算机科学经典教材

数据仓库设计：现代原理与方法

(意大利) Matteo Golfarelli
Stefano Rizzi 著

战晓苏　吴云浩
皮人杰 译

清华大学出版社
北　京

Matteo Golfarelli，Stefano Rizzi
Data Warehouse : teoria e pratica della progettazione, seconda edizione
ISBN: 88-386-6291-6

图书在版编目(CIP)数据

数据仓库设计：现代原理与方法/(意) 戈尔法雪利(Golfarelli，M.) 等著；战晓苏，吴云浩，皮人杰 译. —北京：清华大学出版社，2010.8
(国外计算机科学经典教材)
书名原文：Data Warehouse Design：Modern Principles and Methodologies
ISBN 978-7-302-23074-8

I. 数…　II. ①戈…　②战…　③吴…　④皮…　III. 数据库系统　IV. TP311.13

中国版本图书馆 CIP 数据核字(2010)第 113941 号

责任编辑： 王　军　韩宏志
装帧设计： 孔祥丰
责任校对： 成凤进
责任印制： 王秀菊
出版发行： 清华大学出版社　　**地　　址：** 北京清华大学学研大厦 A 座
http://www.tup.com.cn　　**邮　　编：** 100084
社　总　机： 010-62770175　　**邮　　购：** 010-62786544
投稿与读者服务： 010-62776969，c-service@tup.tsinghua.edu.cn
质　量　反　馈： 010-62772015，zhiliang@tup.tsinghua.edu.cn
印 刷 者： 北京鑫丰华彩印有限公司
装 订 者： 三河市兴旺装订有限公司
经　　销： 全国新华书店
开　　本： 185×260　**印　张：** 26　**字　数：** 600 千字
版　　次： 2010 年 8 月第 1 版　　**印　　次：** 2010 年 8 月第1 次印刷
印　　数： 1～4000
定　　价： 49.80 元

产品编号：033450-01

出 版 说 明

近年来，我国的高等教育特别是计算机学科教育，进行了一系列大的调整和改革，亟需一批门类齐全、具有国际先进水平的计算机经典教材，以适应我国当前计算机科学的教学需要。通过使用国外优秀的计算机科学经典教材，可以了解并吸收国际先进的教学思想和教学方法，使我国的计算机科学教育能够跟上国际计算机教育发展的步伐，从而培养出更多具有国际水准的计算机专业人才，增强我国计算机产业的核心竞争力。为此，我们从国外多家知名的出版机构 Pearson、McGraw-Hill、John Wiley & Sons、Springer、Cengage Learning 等精选、引进了这套“国外计算机科学经典教材”。

作为世界级的图书出版机构，Pearson、McGraw-Hill、John Wiley & Sons、Springer、Cengage Learning 通过与世界级的计算机教育大师携手，每年都为全球的计算机高等教育奉献大量的优秀教材。清华大学出版社和这些世界知名的出版机构长期保持着紧密友好的合作关系，这次引进的“国外计算机科学经典教材”便全是出自上述这些出版机构。同时，为了组织该套教材的出版，我们在国内聘请了一批知名的专家和教授，成立了专门的教材编审委员会。

教材编审委员会的运作从教材的选题阶段即开始启动，各位委员根据国内外高等院校计算机科学及相关专业的现有课程体系，并结合各个专业的培养方向，从上述这些出版机构出版的计算机系列教材中精心挑选针对性强的题材，以保证该套教材的优秀性和领先性，避免出现“低质重复引进”或“高质消化不良”的现象。

为了保证出版质量，我们为该套教材配备了一批经验丰富的编辑、排版、校对人员，制定了更加严格的出版流程。本套教材的译者，全部由对应专业的高校教师或拥有相关经验的 IT 专家担任。每本教材的责编在翻译伊始，就定期不间断地与该书的译者进行交流与反馈。为了尽可能地保留与发扬教材原著的精华，在经过翻译、排版和传统的三审三校之后，我们还请编审委员或相关的专家教授对文稿进行审读，以最大程度地弥补和修正在前面一系列加工过程中对教材造成的误差和瑕疵。

由于时间紧迫和受全体制作人员自身能力所限，该套教材在出版过程中很可能还存在一些遗憾，欢迎广大师生来电来信批评指正。同时，也欢迎读者朋友积极向我们推荐各类优秀的国外计算机教材，共同为我国高等院校计算机教育事业贡献力量。

清华大学出版社

国外计算机科学经典教材

编审委员会

作者简介

Matteo Golfarelli 是意大利博洛尼亚大学计算机科学与技术学院副教授，讲授信息系统、数据库和数据挖掘课程。在全心研究数据仓库设计之前，他担任机器人学和模式识别领域的研究人员。Matteo Golfarelli 毕业于罗得岛州普罗维登斯的布朗大学。他与人合著了超过 60 篇在国际期刊和会议文献上发表的论文，并编著了一本关于信息系统工程的书籍。他参与了多个关于数据仓库设计的研究项目和研究合同。由于他在商业智能领域拥有丰富经验，所以得以在许多博士考试中担任外聘审计员。2008 年以来，他担任了 Business Intelligence Systems Conference 的会议程序副主席，并担任 *International Journal of Data Mining* 和 *Modelling and Management* 的编委。可以通过 matteo.golfarelli@unibo.it 联系他。

Stefano Rizzi 是意大利博洛尼亚大学计算机科学与技术学院教授，讲授高级信息系统和软件工程课程。他在国际期刊和会议文献上已经发表了大约 100 篇关于信息系统、移动机器人系统和模式识别的论文。他参与了多个数据仓库设计的重点研究项目，并参与了所在大学与其他企业之间的合作研究项目。2002 年到 2004 年期间，他担任了博洛尼亚大学管理的“大学数据仓库项目”的学术主管。在 2003 年，他被任命为在新奥尔良市召开的第 6 届 ACM International Workshop on Data Warehousing and OLAP(DOLAP)会议的程序委员会主席。现在他是 DOLAP 指导委员会的成员。2003 年 9 月，他在 International Workshop on Data Warehouse Design and Management(DMDW，2003，德国柏林)会议上发表了一篇题为 Open Problems in Data Warehousing: Eight Years Later 的主题演讲。他是 Tamer Özsu 和 Ling Liu(Springer)主编的 *Encyclopedia of Database Systems* 杂志的数据仓库设计地区编辑。可以通过 stefano.rizzi@unibo.it 联系他。

自从 1997 年以来，两位作者一直致力于数据仓库设计领域的研究。由于在数据仓库的概念设计方面发表了多篇论文，他们在数据仓库领域的国际学术界享有盛誉。他们主持了意大利和其他国家的多家公司、机构和大学的课程和演讲。两位作者都受邀在数据仓库设计方面的国际研讨会上发表主题演讲及参与研究课题的分组讨论。在 1998 年和 2002 年，他们在博洛尼亚大学组织了两次关于数据仓库设计的学术研讨会，与会者都是

企业分析人员和设计人员。在 2001 年，他们在德国海德堡举办的第 17 届 International Conference on Database Engineering 大会上主持了关于数据仓库设计的内容介绍。在 2004 年，他们受邀参加在德国 Dagstuhl 举办的观点研讨会“Data Warehousing at the Crossroads”。这个研讨会旨在为未来几年的数据仓库研究确定新的指导原则。

序　言

数据仓库设计是一个重要的课题，是商业信息的核心，而信息是现代组织的核心，不管组织是一家企业、一家非盈利性组织还是一个政府机构，信息都具有无比重要的价值。就像生物体需要使用信息一样，组织需要使用信息来调节内部操作以及适应外部变化。人员组织使用有关内部运营状况的信息来管理成本及正确地分配资源，使用有关外部世界的信息来管理客户和供应商，以及应对竞争和市场变化。能够很好地利用信息的组织往往能够更加稳定地发展，而不能有效利用信息的组织则在苦苦挣扎。

现在，信息比以往任何时候都更加重要，生成和存储的信息量浩如烟海。我们面临着庞大的信息量。我们居住的模拟世界正在逐渐数字化，而在数字化的世界中，每个事件都可以被记录、归类并存储起来，供以后分析使用。每笔商业交易、每次网站访问、每幅查看的图片、每个打出的电话和花费的每一美元都被数字化并记录下来。在过去，购买就是发生在商店售货员和顾客之间的“模拟交易”，而现在，相同的交易通过互联网或商店的销售点终端系统进行，并将被立即记录下来。

与过去的模拟事件不同，现在的数字事件是多维的。在以前的模拟世界中，购买事件会被记录为“产品 X 以 Y 美元的价格售出”，而在当今的数字世界中，相同的购买事件可能包含另外 10 个属性和维度，比如时间、周几、买主姓名、折扣水平、产品包装、产品促销、颜色、风格、库存位置、销售员姓名和买主同时购买的其他产品的清单。对于每个数字事件，可以同时记录几十个其他属性和维度，以便更好地了解该事件。

与过去的模拟事件不同，现在的数字事件不是作为孤立的事件被记录和读取的，相反，它们彼此关联在一起。商业智能系统可以捆绑相互关联的事件，以便决策者可以在时间流程或者过程流程中看到一系列事件，或者将这些事件看作因果关系，甚至看作预测计算。在上面的购买示例中，如果是模拟世界，那么所有的购买事件将被汇总到一起，以计算出一天内每件产品的总销售额；而如果是数字世界，公司希望跟踪每个客户的购买模式和产品偏好，以及了解促销和产品放置位置的效果，以便能够向每个客户销售更多产品，并预测将来的库存需求和需求水平。

遗憾的是，数字世界并不能自行组织所有生成的数据。事实上，信息爆炸已经使得每个系统在捕获、组织和标记数据的方式上各有不同。基于在仓库数据模型中编码的整体业务模型，数据仓库消除了这种差别。

数字化世界最重要的意义在于组织将捕获更多的事件，将更多的维度与每个事件关联起来，以及使用越来越复杂、富有创意并且无法预测的方式将这些事件彼此关联起来。而这又意味着更多的数据、数据项之间的更多关联和更复杂的数据提取。现在，包含 50TB 数据、跨越很多维度并包含数百个主要指标的数据仓库十分常见。所有这些特征要求数据仓库提供非常高的性能(容量大)并且在使用和演变方面非常灵活。这是设计现代数据仓库时面临的真正挑战——在具有高性能的同时具有很高的灵活性和卓越的功能。

数据仓库设计是一个工程课题。与所有工程课题一样，高性能几乎总是以降低灵活性、减少功能为代价。只要愿意放弃一些将来的灵活性，实现高性能的系统会简单得多。同理，工程师可以创建高度灵活、功能很强的系统，但是这将以降低系统性能为代价。工程学就是要在高性能和高灵活性、强大的功能之间找出一种可行的平衡方案。这仍然是数据仓库工程师面临的最大挑战。多维数据库(例如立方体数据库)可以提供很高的性能，但是只能应用到作为立方体模型的一部分创建的计算和视图中。相反，面向对象的数据库具有几乎不限制数据检索的灵活性并允许平稳地演进数据模型，却没有很高的性能或可扩展性。现代关系数据库可以在大规模环境中提供很高的性能，并且在与能够创建复杂 SQL 的强大的商业智能(BI)结合使用时，可在数据提取方面提供很大的灵活性。

数据仓库设计是一个工程课题，许多人却希望它不是。他们希望数据模型并不复杂，希望聚合策略和索引方案很简单，希望不必清晰定义所有的数据关联，希望所有的 BI 工具能从逻辑上建立任何数据结构的模型，希望全部数据结构的性能都很高。但是些都是凭空臆想。只有付出艰苦努力，设计、填充和演化数据仓库，才能得到灵活性和性能都很好的数据仓库。

本书提供了数据仓库工程师所需的一个工程框架、技术和工具，帮助他们提供成功的数据仓库。对于想要利用新一代数字技术的全部功能的下一代数据仓库工程师来说，这是一本重要的典籍。

Mark Stephen LaRow
MicroStrategy 产品管理副总裁

前　言

数据仓库是历史的、集成的和一致的数据的储存库。通过使用数据仓库的各种工具，公司管理层可以提取可靠的信息，并将其用于支持决策过程。数据仓库设计涉及到从企业信息系统提取相关数据、转换数据、集成数据、清除缺陷和不一致数据以及将数据存储到数据仓库的过程，并允许终端用户访问数据，以便他们执行复杂的数据分析和预测查询。

当前，多家大中型组织已经建立了很出色的数据仓库系统，并且在为终端用户提供访问重要信息的简便性方面，数据仓库系统扮演了一种具有战略意义的角色。数据仓库设计最早诞生在企业界，用于应对用户不断增长的需求。一开始学术界是忽略这种现象的，并认为它只是一个技术问题，直到他们认识到数据仓库设计的复杂性和面临的挑战，这种看法才得到改变。实际上，数据仓库设计这个主题已经成为国际学术会议的主要部分。自此之后，许多会议和研讨会都专门研究数据仓库设计。

在讲授了超过 10 年的数据库设计课程以后，我们开始研究数据仓库设计领域。我们的目的是找出是否应该丢弃为关系系统开发和测试的设计技术，或者是否也可以把它们部分应用于数据仓库设计。我们很快认识到，一方面，与用户需求的概念建模有关的基础问题都被忽视了；另一方面，杰出的研究人员已经撰写了大量关于如何优化数据仓库性能的文章，这些对这个领域起到了奠基性的作用。

我们两人都认为，在创建软件系统时，采用基于可靠的软件工程原理的系统方法十分重要。特别是，在开发数据库时，需要进行精确的概念建模。常见的数据仓库设计实践不具备合理的系统方法和概念模型。为了填补这个空白，我们努力开发出了一个全面的、有充分依据的设计方法，来帮助设计人员正确地处理数据仓库项目的各个阶段。我们的方法以使用显示数据仓库应用程序的具体特征的概念模型为中心，同时仍然基于使设计人员可以设计和创建更好项目的简易方法。我们提出的方法和概念模型的有效性已经在意大利的多家软件公司和顾问的多个项目中得到印证，也得到了专家的首肯。

本书的主要目的有两个：一是总结大量的设计技术，并使之成为系统化和全面性的

方法架构的一部分，供设计人员参考；二是适合学生阅读，因为虽然一些大学已经开设了数据仓库设计类的课程，但是相关用书太少。本书收集并系统化最新的一些关于数据仓库设计的重要研究成果，并总结了在多家公司应用我们的设计技术后积累的经验。

我们的学术背景和专业经验使我们不只能够在讲解数据仓库设计时兼顾实用方法和正式理论，还使得我们能够引起人们对企业专用的应用程序和尖端研究成果的关注。我们在书中提供了大量示例，并且提供了一个现实案例研究，以便最大限度地强化这些信息的教学效果。

我们的方法在企业界的广泛应用促使我们提供了一个计算机辅助软件工程(Computer-Aided Software Engineering，CASE)工具，用以支持数据仓库设计的核心阶段和相关项目文档。对此工具感兴趣的读者可以访问网址 www.qbx-tool.com。

读者对象

本书适用于对了解决策支持系统领域感兴趣的数据仓库设计人员、公司和机构，也适用于学习信息系统和数据库方面的高级课程的学生。读者应该已经熟悉关系模型和实体-关系模型的基本知识，以及信息系统和信息系统设计方面的知识，否则无法从本书讨论的概念中收到最圆满的学习效果。

本书结构

第 1 章介绍了数据仓库设计领域使用的基本定义。本章回顾了可以应用的许多功能体系结构，并详细讨论了一些要点，以帮助读者理解整个过程。这些要点为：使用数据源为数据仓库提供数据，作为数据仓库的构建基础的多维模型，以及用户在访问数据仓库信息时具有的几个重要选项。本章简要讨论了实现数据仓库的两种主要方法：基于关系模型的方法和使用专门的多维解决方案的方法。后面的章节中重点讨论第 1 种方法，因为它远比第 2 种方法受欢迎。

第 2 章描述了数据仓库系统的生命周期，并提出了一种系统的方法来设计它们。本章为后面的章节提供了一个参考点。本章描述了生命周期的 7 个阶段，后面的 7 章分别讨论每个阶段。本章还建议数据仓库设计由 3 个主要阶段(概念设计、逻辑设计和物理设计)组成，这样就可以成功地开发出传统的信息系统。此外，还介绍了 3 个不同设计场景：数据驱动的场景、需求驱动的场景和混合场景。提出这些场景是考虑到现实公司的情况各异，并且是为了使我们的系统方法更灵活。

第 3 章描述了为数据仓库提供数据的数据源分析和协调阶段。这个阶段对于确保获取的信息满足最高质量标准要求至关重要。本章深入讨论了集成多个异构数据源时涉及的活动。

第 4 章讨论了需求分析阶段。只有认真执行这个阶段才能确保构建的系统能够满足终端用户的需求和期望。我们介绍了两种方法：基于简单词汇表的非正式方法和用于需求驱动设计或混合设计的正式方法。

第 5 章提出了作为系统方法基础的概念模型——维度事实模型(Dimensional Fact Model，DFM)。我们逐渐介绍这个模型的结构，以帮助读者深入了解从建模基础到现实应用程序所需的表示细节等各种知识。

第 6 章介绍概念设计，并展示如何利用数据源文档和用户需求来为数据仓库创建概念模式。

第 7 章描述了产生概念模式中的工作负荷表达式的阶段。这个主题对于后面的逻辑和物理设计阶段(向设计人员展示如何优化性能)十分关键。本章还讨论了与数据卷定义有关的重要主题。

第 8 章介绍了关系数据仓库中最流行的逻辑建模方法。本章特别关注著名的星型架构的描述以及冗余视图，以便提高数据仓库的性能。

第 9 章讨论了基于概念模式的逻辑设计，以及设计人员可以使用的逻辑优化策略和技术——最重要的是实体化视图。本章基于关系架构为每个 DFM 构造提供可能的实现。

第 10 章描述了如何为数据仓库提供数据。在这个阶段中从数据源提取数据、转换数据、清除数据，然后填充数据仓库。

第 11 章列出并分析了可以在数据仓库系统中使用的主要索引类别和主要的联接算法。

第 12 章描述了物理设计。本章首先讨论如何选择最合适的索引，然后讨论了与物理设计有关的其他主题，比如增加规模和分配量。

第 13 章向读者展示如何创建有效完整的项目文档，以便向设计人员提供参考，并帮助维护将来的系统。

第 14 章分析了一个基于现实经验的案例，并示范如何按照前面章节中描述的系统方法的基本步骤解决问题。

第 15 章简要介绍了商业智能，这是一个覆盖面很广的领域，植根于数据仓库系统，并受益于用来满足业务执行者需要的高级解决方案和体系结构。

阅读建议

针对本书的 3 类主要读者，我们提供了 3 种阅读方法，以满足各类读者的需要，并且使大多数读者享受到阅读的乐趣。

初学者

本书为初学者提供了主要设计问题的全面概述，具体内容详见以下章节：

第 1 章：全部内容
第 2 章：2.1、2.2、2.3 和 2.4 节
第 3 章：引言
第 4 章：引言
第 5 章：5.1、5.2.1、5.2.3、5.2.4、5.2.6 和 5.2.9 节
第 6 章：引言和 6.5 节

第 7 章：7.1.1 和 7.1.5 节
第 8 章：8.1、8.2 和 8.3 节
第 9 章：引言
第 10 章：引言
第 12 章：12.2 节
第 13 章：全部内容
第 15 章：全部内容

设计人员

决策支持系统的设计人员可将本书用作一个关于数据仓库设计的方法手册，以及针对具体问题的详细解决方案的参考。

第 1 章：全部内容
第 2 章：全部内容
第 3 章：全部内容
第 4 章：全部内容
第 5 章：5.1、5.2、5.3、5.4 和 5.5 节
第 6 章：6.1、6.2、6.4 和 6.5 节
第 7 章：全部内容
第 8 章：全部内容
第 9 章：9.1、9.2.1 和 9.3.2 节
第 10 章：全部内容
第 11 章：11.1、11.2.1、11.4 和 11.6 节
第 12 章：全部内容
第 13 章：全部内容
第 14 章：全部内容

学生

本书也为学习高级信息系统课程的学生提供了关于理论和应用问题的全面介绍，此外也可以满足以实用为主的大学课程的需要。

第 1 章：全部内容
第 2 章：2.2、2.3 和 2.4 节
第 3 章：全部内容
第 4 章：4.2 节
第 5 章：全部内容
第 6 章：全部内容
第 7 章：全部内容
第 8 章：全部内容

第 9 章：全部内容
第 10 章：引言
第 11 章：全部内容
第 12 章：12.1 和 12.2 节
第 14 章：全部内容
第 15 章：全部内容

目　录