Broadview
WWW.BROADVIEW.COM.CN

数据仓库与数据挖掘
技术应用丛书

The Data Warehouse Lifecycle Toolkit : Expert Methods for Designing, Developing, and Deploying Data Warehouses

数据仓库生命周期工具箱：设计、开发和部署数据仓库的专家方法

[美] Ralph Kimball Laura Reeves
Margy Ross Warren Thornthwaite 著

肖明 王永红 等译

含光盘

WILEY
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

数据仓库生命周期工具箱

——设计、开发和部署数据仓库的专家方法

The Data Warehouse Lifecycle Toolkit:
Expert Methods for Designing, Developing, and
Deploying Data Warehouses

[美] Ralph Kimball　Laura Reeves
Margy Ross　Warren Thornthwaite　著

肖明　王永红　等译

電子工業出版社

Publishing House of Electronics Industry

北京·BEIJING

内 容 简 介

本书是著名数据仓库畅销书作者 Ralph Kimball 的著名作品，在世界各地畅销不衰。这是目前惟一一本从技术和管理两个角度介绍了使数据仓库项目获得成功所必备的各种知识和经验教训的专著，这些内容都是作者自 1982 年以来在从事数以百计的数据仓库安装和咨询任务过程中不断积累总结出来的。书末的两个附录中提供了大量的框架、任务、模板以及生动详实的样例（具体内容见本书配套光盘），所有这些都使本书别具一格。全书主题广泛，思想深刻，内容详尽，图文并茂。

本书不仅是现代信息系统开发人员的重要指南，而且是所有面向数据仓库项目的设计、开发、管理和咨询人员的高级参谋，并且适合信息管理与信息系统、计算机应用、电子商务等专业的高校师生作为教学参考用书，还可供从事传统数据库系统工作的技术人员参考阅读。

图书在版编目（CIP）数据

数据仓库生命周期工具箱——设计、开发和部署数据仓库的专家方法/（美）金博尔（Kimball,R.）等著；肖明等译. —北京：电子工业出版社，2004.1（数据仓库与数据挖掘技术应用丛书）

书名原文：The Data Warehouse Lifecycle Toolkit：Expert Methods for Designing, Developing, and Deploying Data Warehouses

ISBN 7-5053-9192-5

Ⅰ.数…　Ⅱ.①金…②肖…　Ⅲ.数据库系统　Ⅳ. TP311.13

中国版本图书馆 CIP 数据核字（2003）第 086309 号

责任编辑：张立红　zlh@phei.com.cn

印　　刷：北京增富印刷有限公司

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：46.5　字数：814

印　　次：2004 年 1 月第 1 次印刷

印　　数：5 000 定价：86.00 元（含光盘）

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。

出 版 说 明

如果没有对海量数据进行科学分析的能力，沃尔玛的老板再精明，也绝对想不到“啤酒与尿布”这两个风马牛不相及的东西之间还有着千丝万缕的联系。而将它们放在一起，竟然增加了啤酒销量，可见数据分析的巨大威力。

信息系统数年中收集了海量数据，且数据还正以指数级增长，企业迫切地需要高效、精确、科学地分析数据，以找出其背后的寓意，进而了解企业的经营状况和外部环境，做出科学的决断，在现代激烈的竞争中胜出。所以，如何将数据点石成金，更是摆在我们面前很现实也很诱人的一个问题。

现在，很多人已经意识到数据中潜在的大量商机，并踏踏实实地进行着从数据中沙里淘金的工作。特别是在信息化的大潮中，上至政府，下到企业，从银行到电信，再到网站、超市，人们都希望用数据分析这根魔杖赢得先机。与此同时，人们也在期盼着相关书籍，以便工作中学习参考。在广泛征询专家和用户的基础上，秉着选题全面、内容经典、译者严谨的原则，我们适时地推出了这套《数据仓库与数据挖掘技术应用丛书》，以飨读者。本丛书有如下几本：

- 数据仓库基础
- OLAP 解决方案：多维信息系统的构建技术
- 数据仓库工具箱：维度建模的完全指南（第二版）
- 数据仓库生命周期工具箱：设计、开发和部署数据仓库的专家方法
- 数据仓库及其在电信领域中的应用
- 疑难数据仓库专家解决方案
- IBM 数据仓库和商业智能工具
- 可视化数据挖掘：数据可视化和挖掘的技术和工具
- 点击流数据仓库
- Web 数据挖掘：将客户数据转化为客户价值
- 企业信息工厂
- 机器学习与数据挖掘：方法和应用

本丛书既包括商业智能（BI）的基础——数据仓库（DW），也包括数据仓库上的两类不同目的的数据增值操作——联机分析处理（OLAP）和数据挖掘（DM）；既覆盖基础理论，如数据仓库基础，又提供不同领域的解决方案，如数据仓库在电信、银行、保险等领域的应用。

本丛书来自国外数据库领域一些著名作者的畅销书，以及国内第一线实施

者的精心总结。如一直位居 AMAZON 畅销书榜的数据仓库领域的畅销书作家 Ralph Kimball 的《数据仓库工具箱：维度建模的完全指南（第二版）》、《数据仓库生命周期工具箱：设计、开发和部署数据仓库的专家方法》，数据仓库之父 William H.Inmon 的《企业信息工厂》（Corporate Information Factory）等。

丛书的译者均来自工作在该领域一线的人员，既有该领域的理论和实践经验，又具备中英文翻译的功底。且多位译者先前均已读过原著，所以，自感翻译的过程不再枯燥，而是情趣盎然，乐在其中。

出版高品位、高品质的图书是博文视点的努力目标。希望您对我们的工作多提宝贵意见。您的意见是我们创造精品的动力源泉。

如果您希望将您的工作经验感悟等总结成书，我们将为您提供一流的服务，共创精品图书。

我们的联系方式如下：

地址：北京复兴路 47 号天行建商务大厦 604

邮编：100036

电话：010-51922832，68216158

传真：010-51922823

E-mail：jsj@phei.com.cn；zsh@phei.com.cn

博文视点资讯有限公司

2003 年 10 月

译 者 序

对于今天的业界人士来说，“数据仓库”早就是耳熟能详的专业术语了。但是，对于我来说，与“数据仓库”的第一次亲密接触是在 1993 年的一次专业会议上，对它真正有所了解却是在 5 年之后。1998 年，我在中国科学院攻读博士学位期间，参加了由导师负责的一个“数据仓库”科研项目，并且选择“文本挖掘”作为我的主攻科研方向。在从事这些科研活动的过程中，Ralph Kimball 的名字不断闪现在有关“数据仓库”的各种专业文献中。于是，我决定探询其中的缘由。后来，我陆续搜集到 Ralph Kimball 的许多资料，知道他是美国乃至全世界范围内非常著名的“数据仓库”专家。如今，Ralph Kimball 的名声更响了，英文 YAHOO 索引库中目前就存有 50 800 多条 Ralph Kimball 的索引信息。

1996 年 2 月，著名的 John Wiley & Sons 出版公司出版了 Ralph Kimball 的专著《数据仓库工具箱》（The Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data Warehouses），该书由被誉为“数据仓库之父”的 W. H. Inmon 作序，获得了极大的成功，来自读者和出版商的评价都非常高。后来，Ralph Kimball 又和三位富有实践经验的合著者一起对该书重新进行了修订。于是，就有了这本《数据仓库生命周期工具箱》（The Data Warehouse Lifecycle Toolkit : Expert Methods for Designing, Developing, and Deploying Data Warehouses）。

《数据仓库生命周期工具箱》刚一问世，好评如潮。在 Barnesand -noble.com 公司网站上，有读者评价说：“尽管我在项目管理和数据建模方面积累了多年的实践经验，但我认为本书才是解决有关项目管理和数据仓库生命周期方面问题的最好参考信息源。” 还有读者认为“本书要比第一版更加物有所值，它探讨了有关数据仓库的设计、实现、运行、维护和管理等方面的几乎所有问题。”

作为本书的主要译者和审校者，我觉得本书的最大特色可以归纳为：“主题广泛，思想深刻，内容详尽”。本书深入探讨了业务维生命周期方法和数据仓库总线结构这两个主题，介绍了使数据仓库项目获得成功时必备的各种知识和经验，还提供了大量的框架、任务、模板以及分步骤介绍的各种解释和样例，所有这些都使得本书在同类书中别具一格。本书特别适合面向数据仓库的开发人员、管理人员、设计人员、数据管理员以及其他相关人员。对于计算机、信息管理以及其他相关专业的本科生和研究生也具有重要参考

价值。

参加本书翻译的人员主要由来自北京师范大学、北京大学、联想集团等单位从事数据仓库和数据挖掘研究的教学科研人员组成。他们是：肖明、王永红、张敏、郭宇峰、宋雯雯、于超和唐琳琳。全书由北京师范大学肖明副教授审校，所有插图由王永红负责绘制和修改。

由于时间紧迫，加上译者水平有限，目前有关数据仓库的术语尚不规范，所以书中难免有错误和疏漏之处，敬请广大读者不吝指教。如有任何批评意见和建议，请与译者联系 ming_xiao02@sohu.com。

肖明

译者介绍

肖明，北京师范大学信息技术与信息管理系副教授，博士。主要研究方向为：互联网文本挖掘、多媒体信息处理、信息系统检索、计算机网络。已发表 40 多篇论文、译文 100 多万字，参加编译书籍 9 部，独立编写教材 1 部，参加包括国家 863 计划项目、国家教委教改项目、国家科委攻关项目等在内的 4 项国家级科研项目，主持国家青年基金项目两项（在研）。目前从事教学的课程主要有：计算机组成原理与体系结构、信息资源管理、中文信息处理、计算机网络等。代表性著作主要有：《电脑文化简明教程》、《信息系统分析与设计》（以上两本教材为教育部审订面向 21 世纪教材）、《数字图书馆：概念与技术实现》、《信息资源管理》等。

前　言

本书的主要读者对象是数据仓库的设计人员、管理人员及其所有者。他们都是在信息系统组织中工作的。无论他们的具体头衔是什么，都会深感创建和维护数据仓库（或是数据仓库的某些部分，即我们通常所说的数据集市）的责任重大。我们通过这本《数据仓库生命周期工具箱》提供了一份领域指南以及一整套工具，主要用于设计、开发和部署大型组织机构的数据仓库和数据集市。

我们试图使本书内容非常具体化和可操作，因此，它与其他数据仓库图书完全不同。我们在本书中描述了同完整数据仓库范围完全一致的框架，从关于开发和部署数据仓库的所有详细步骤，到用于规划其下一阶段的最终步骤，可以说贯穿了数据仓库的整个生命周期。

数据仓库市场显然已走出了其幼年时期。到本书写作时为止，仅美国就安装有一千多个数据仓库，它们都在发挥着作用。目前，许多数据仓库所有者全都按本书中介绍的“生命周期”观点看待其数据仓库。也许从生命周期观点中得到的最大收获就是：意识到各个数据仓库都在不断发展，处于动态变化之中，各种新的业务要求不断产生。新的管理人员和主管将一些不可预知的要求提交给数据仓库，还可以获得各种新的数据源。或者说，数据仓库至少需要尽量随着组织环境的变化而不断取得进展。稳定型组织会要求数据仓库取得适度进展，而变动较大的动态型组织则可能会使数据仓库任务变得富有挑战性。

考虑到数据仓库目前所具有的动态发展特性，我们必须对若干年前基于原始、理想化、静态的观点所提出的数据仓库期望值和相关技术进行调整。我们必须采用那些具有灵活性、可修改的各种技术，还必须同时扮演 DBA 和 MBA 双重角色。我们还需要见机行事地将一些小块数据（比如数据集市）连成大块数据（即数据仓库）。同时，还要求针对数据仓库所做的变化都必须是完美的。完美变化意味着以前的数据和各种应用仍然有效。

本书深入探讨了两个主题。第一个主题是业务维生命周期（Business Dimensional Lifecycle）方法。业务维生命周期始于业务需求，并创建了一系列具有可理解性、高性能的数据集市。这些数据集市全都是星型维度模型。

第二个主题是数据仓库总线结构。本书中介绍了如何创建一系列数据集市，使读者能够及时创建一个完整的数据仓库。在发布第一个数据集市之前，利用该方法就可以依据需求轻松地创建一个无所不包的、集中式数据仓库。

本书中涵盖了上述这些观点，提供了能够帮助读者完成作业任务的各种有用的技巧和工具，并且通过这种方式来介绍我们所积累的主要观点和价值

观念。它们都是我们自 1982 年以来在从事数以百计的数据仓库安装和咨询任务过程中不断积累起来的。

本书特点及适用的读者

本书的主要读者对象应该是那些从事数据仓库的创建和管理工作的设计人员或者管理人员。本书还包含了一些介绍性材料，这些材料对于与数据仓库相关的信息系统专业人员来说也许很有用。熟悉 Ralph Kimball 所著的《数据仓库工具箱》（Wiley 公司 1996 年出版）一书以后，就会了解数据仓库方面的适当背景知识。本书是建立在前一本书（《数据仓库工具箱》）的“工具箱”概念基础之上，但本书能提供更深入、更先进的数据仓库开发方法。

此外，通过设计和开发一个真实的数据仓库，就能够积累一些数据仓库经验，并形成自己的观点，这是最好的知识背景。没有任何东西可以替代在开发一个有效的数据仓库时所承担的责任。我们都曾有过感到羞辱的经历，那就是将数据仓库介绍给一群要求过分的最终用户的时候。通常令人难以接受的事实是，大多数最终用户的工作与技术毫不相关，他们甚至可能不是特别喜欢技术。但是，如果我们的技术易于使用，并且能为用户提供确有实效的使用价值，最终用户还是会使用我们的技术的。

本书略微偏向技术。其中有关数据仓库的设计技术和结构等方面的讨论，将介绍一些未曾遇见过的术语。我们对本书进行了精心梳理，以确保那些倾向于技术方面的主题都是我们认为读者必须懂得的内容，我们不打算因内容本身方面的缘故而陷入细节上的困扰。例如，对篇幅较长的、有关数据仓库安全性的章节内容的处理。有关安全性的讨论中我们尽量避免描述安全技术的精微细节，并注意不占据太大的篇幅。同时确保读者在承担某种安全责任时，能了解足够多的安全性主题。

如何有效使用本书

我们建议读者在了解感兴趣的章节前，将本书通读一遍，以便获得完整的业务维生命周期知识。各种经验和意见可能会帮助形成这方面的个人观点框架。例如，在读完第 2 章后，也许会明白在创建数据仓库时必须抓住三条平行线索，即技术结构、数据结构和应用结构。在各章开始部分的那个图形中的“You Are Here”处（译者注：即图中加阴影的部分）展示了这三条线索。尽管这三条线索之间显然会相互影响，但它们可以按平行方式或者异步方式发展。

由于图书的内容按线性方式进行编排，所以，书中介绍的业务维护生命

周期的所有步骤，就像是按某种固定次序发生的那样。因此，在读完本书以后，就能够想像出这些步骤在现实世界中具有更现实、更复杂的各种关系。

本书中融合了许多实用技巧，为了便于读者轻松阅读，我们采用了如下的一些标识。

看到形如电灯泡的这种图标，就能找到各种技巧。

建立数据仓库过程特定部分的快照、项目计划任务和电子表格，请按以下标识分别进行查找。

这个标识在一些章节后面。它收集了有关数据仓库生命周期各个阶段所起的主要作用，其指示标记是由三根钥匙构成的一个图标。

这个标识在一些章节后面。它收集了有关各个处理步骤预计要考虑的各种事项，其指示标记是一个闹钟图标。

这个标识在一些章节后面。它收集了有关各种支持性模板的一份清单，其指示标记是一个 CD-ROM 图标。打开 CD-ROM 就可以获得相应的空白模板，然后按提示内容去使用它。

在全书中到处都在谈论着各式各样的规划所需帮助、清单以及模板。由于它们会对读者有所帮助，所以，建议使用 CD-ROM 中提供的各种样例。或许读者已经形成了自己的独特风格，或者已经拥有不同于我们的规划框架。无论是哪一种情况，我们所做的目的都是帮助读者尽快上路。

本书中给出的规划所需帮助仅提供了一个中等程度而不是详细程度的结构。由于数据仓库的实现是一项巨大的工程，所以，从事这方面工作的任何人都必须是非常好的管理人员。而好的管理人员都应该懂得如何平衡项目管理方法论以及人员和任务管理中的人员和逻辑问题等之间的关系。所以，建议使用本书中给出的结构或者读者自己创建的结构，但不要过分依赖它们。真正要做的工作是首先判断出组织中哪些是必须要做的重要事情，接着与其他人一起工作并完成它。

还可以将本书分成一条基本线索和一条“研究生”线索。本书目录中已清楚地标明其中有三章属于研究生课程。因此，在第一次阅读本书时（特别

是当大部分材料都是新内容时更是如此），应该跳过那些标明为研究生课程的章节内容，只要了解其大致内容即可。接着，在对整个业务维生命周期掌握得更加得心应手以后，就会发现那些标明为研究生课程的章节内容都是非常有价值的。这些章节内容所介绍的都是这三个领域中的最新思想。

当项目进入到特定阶段时，应该返回到相应的章节，并且非常仔细地阅读其内容。这也正是为什么本书被命名为《数据仓库生命周期工具箱》的真正原因。

各章写作目的

第 1 章　数据仓库的基本组成

在本书撰写时，数据仓库方面许多含义不清的术语四处泛滥，甚至连数据仓库这个概念也失去了其准确含义。一些人甚至尝试将数据仓库定义为一种不可查询的数据资源。本章试图解决术语方面的所有争端，本书中统一使用术语的某种特定含义。本章中以统一的方式简要地定义了数据仓库中使用的一些重要术语。这或许有点儿像在打算下一盘棋之前必须研究所有的棋子及其用法。本书中所给出的术语定义都非常接近于它们的主流定义。

第一部分　项目管理与需求

第 2 章　业务维生命周期

本章从非常高的高度对整个业务维生命周期进行了定义，还简要讨论了其中的每一个步骤，并给出了对生命周期的整体看法。

第 3 章　项目规划与管理

本章对项目进行了定义，探讨了如何在考虑组织环境因素的同时设定项目的范围。此外，还广泛谈论了各种项目中的角色和责任。但大可不必一一调查所有项目的角色，只需要代之以任何可以想像得到的具体项目即可。因此，本章主要是写给管理人员看的。

第 4 章　收集项目需求

收集有关业务和数据的需求是整个数据仓库项目的基础，或者至少应该这样做。收集项目需求需要一定的技巧，并且它是信息系统组织中最常见的一项活动内容。本章提供了能轻松完成该工作的各种技术，但读者不必在该步骤上花费太多的时间。

第二部分　数据设计

第 5 章　维度建模的第 1 课

本章开头部分积极讨论了维度建模的价值。应该理解本章介绍该方法的深度。在过去的 15 年中，当我们完成了数以百计的数据仓库设计和安装任务以后，我们认为该方法是能够实现易理解性和性能这两大目标的惟一方法。接着，我们展示了如何将各种多维模型组合到某种一致性模型中的重要秘密。这个秘密就是所谓的一致性维度和一致性事实。我们将该方法称做数据仓库总线结构。计算机中有一个重要部件（即计算机总线），用户可以将所有东西连接到该总线上。同样地，数据仓库中也有一个重要部件，我们称它为数据仓库总线，也可将所有东西连接到它上面。本章的剩余部分全面介绍了有关数据仓库维度建模的知识，这个介绍可以看做 Ralph Kimball 先生所著的《数据仓库工具箱》一书中所论及主题的附录。

第 6 章　维度建模的研究生课程

本章收集了我们所能想到的维度建模方面最艰难的各种情形。其中的大多数例子来自特定的业务情形，比如，如何处理一些奇怪的客户。

第 7 章　多维模型的创建

本章需要解决的是如何为组织创建一个合适的模型。首先，需要建立一个有关数据集市和维度的矩阵。接着，可以按第 5 章中所描述的各种技术为每一个数据集市设计各种事实表。本章的后半部分描述了各种现实的管理问

题，这些问题都是在应用上述方法以及创建各个数据集市所必需的所有维度模型时遇到的。

第三部分　数据仓库结构

第 8 章　数据仓库结构介绍

本章按照中等详细程度介绍了数据仓库技术结构的全部部件，描述了其中的全部情景。本部分剩下的 5 章探讨了特定领域的细节情况。这方面的讨论可以细分成数据结构、应用结构以及体系结构等部分。在遵循第 5 章中提出的数据仓库总线结构以后，就能够每次创建一个数据集市，并且最终能够得到一个灵活的、统一的完整数据仓库。但是，这并不意味着很容易就能够完成这件事。

第 9 章　后台技术结构

本章介绍了后台的各种系统部件，包括源系统、报告实例、数据登台区、基础级数据仓库和业务处理数据集市。本章中将介绍有关操作型数据存储（ODS）的情况。还会讨论后台中必须提供的所有服务，利用它们可将数据装载到数据集市呈现服务器中。

第 10 章　前台结构

前台就是执行发行操作的地方。应该使数据可以获得，并且提供用来满足不同用户需求的一组工具。本章还提供了在前台中必须支持的许多需求的全面性观点。

第 11 章　体系结构与元数据

体系结构是用来将数据仓库连成一个整体的。本章中包含了体系结构的具体细节情况。在讨论细节情况时，考虑的是每一位数据仓库设计人员和管理人员都必须了解的硬件、软件、通信等方面的知识，特别是元数据知识。

第 12 章　有关互联网和安全性方面的研究生课程

尽管互联网已经对数据仓库管理人员的生活产生了极其巨大的潜在影响，但许多数据仓库管理人员不是没有认识到互联网对他们的真实影响，就是避免讨论这方面的问题。本章将展示基于互联网的数据仓库及其安全性等方面问题的现状情况，还提供了用来保护数据仓库安装过程安全的一份行动清单。贯穿本章的各种样例都倾向于揭示数据仓库拥有者必须面对的各种挑战和内幕。

第 13 章　创建结构计划与产品选择

本章假设读者是一位软件、硬件、体系结构等方面的专家，正准备为组织制定一份具体的结构计划，还负责选择各种具体产品。本章中讨论了产品选择过程以及组合产品策略。但需要记住的是，本书中并不打算讨论某些具体销售商的产品平台情况。

第四部分　数据仓库实施

第 14 章　有关聚集的研究生课程

聚集是指创建的预存储概要，主要用于提升数据仓库系统的性能。本章深入探讨了聚集的结构、聚集应用的场合、如何使用聚集以及如何管理聚集等内容。假如其他系统是按数据仓库总线结构进行建造的，则聚集就是用来提升大型数据仓库系统性能的一个性价比最高的途径。

第 15 章　完成物理设计

尽管不了解读者会选择哪一种数据库管理系统和硬件结构，但我们仍建议读者了解这方面的许多重要思想。本章中讨论了物理数据结构、索引策略等内容，特别是讨论了用于数据仓库的各种专业数据库以及 RAID 存储策略。

第 16 章　数据登台

一旦安排好了各种主要的系统，则接下来的就是最艰巨、风险最大的处理步骤，即需要从传统系统中取出数据，并将该数据装载到数据集市数据库管理系统中。数据登台区是用来临时存放要进行净化和转换的传统数据的中转地。本章详细讨论了数据登台区中可能会发生哪些情况以及不应该发生哪些情况。

第 17 章　创建最终用户应用

在数据终于被装载到数据库管理系统以后，还必须安排如何在用户桌面上进行“软着陆”方面的事情。最终用户应用是指各种查询工具、报告写作程序和数据挖掘系统，其主要功能是从数据库管理系统中提取数据并实现一些有用的功能。本章描述的是用于起步阶段的一组最终用户应用，它们都是你在数据集市实施的起始阶段所必须提供的各种应用。

第五部分　数据仓库的部署与增长

第 18 章　规划部署

在一切准备就绪以后，应该暂时抛开该系统，并且像商业软件销售商那样采取行动。必须做的事情包括：安装软件、培训用户、收集错误报告、征求反馈意见和响应各种新需求。还必须小心翼翼地制定各种计划，以便能按设定的期望值交付该系统。

第 19 章　数据仓库的维护与增长

最后，当整个数据集市建立起来并运转以后，还必须回过头来再做一遍。但正如先前所说过的那样，与其说数据仓库是一个过程，还不如说数据仓库是一个项目。当本章能为读者留下这样一种有价值的最后印象时（即“所做的事情永远不会完毕！”），说明它最适合做本书的结束部分。

各种支持工具

- 附录 A

附录 A 中总结了在业务维生命周期的某个地方或者用某种格式需要用到的整个项目计划。其中列出了全部项目的任务和角色。

- 附录 B

附录 B 是本书附带的 CD-ROM 的一份内容导游图。还遍历了如何使用数据仓库总线结构样例设计。

- CD-ROM

本书附带的 CD-ROM 中包含了大量实用的检查清单、模板以及可用于数据仓库开发的各种表格，其中还包括用来描述数据仓库总线结构的样例设计。

数据仓库的目标

组织中最重要的资产就是它所拥有的信息。这种信息资产通常保存成以下两种形式，即操作型记录系统和数据仓库。简言之，操作型记录系统是指存放数据的地方，而数据仓库是指能从其中取出数据的地方。《数据仓库工具箱》一书中曾经详细描述过这种二分法。在本书写作时，似乎没有必要再让人确信整个世界的确只存在两类系统，或者经常只存在两类系统。目前能被广泛接受的一种观点就是，数据仓库要比操作型记录系统具有更多的需求、客户、结构和节奏。

最后还需要暂时撇开数据仓库的实现和建模等方面的细节情况，并且要牢记到底什么是数据仓库的基本目标。数据仓库具有以下特点。

- **使组织信息变得可存取**

数据仓库的内容都是可理解、可浏览的，数据仓库的存取表现为快速的性能。这些方面的要求既无边界，也没有明确的限制。“可理解”意味着需要为其内容加上正确的标签，使之显而易见。“可浏览”意味着需要认识到数据仓库的终极目的地是用户屏幕，用户只需要单击一下就可以浏览相关内容。“快速的性能”意味着零等待时间。其他事情都意味着某种折中，所以必须在某些方面有所改进。

- **使组织信息具有一致性**

来自组织中某一部分的信息必须与另一部分的信息相匹配。当组织中的两种指标方法名称相同时，它们肯定是指同一件事。反之，当它们不是指同一件事时，其标识也应该不一样。信息的一致性意味着信息的高质量，还意味着所有信息都是可以证明的完整信息。其他事情都意味着某种折中，所以

必须在某些方面有所改进。

- **它是一种自适应的、有弹性的信息源**

数据仓库被设计用于持续变化环境。当提交有关数据仓库的各种新问题时，现有的数据和技术都不会发生变化或者遭到破坏。当新数据被添加到数据仓库时，现有的数据和技术都不会发生变化或者遭到破坏。由于多个数据集市可以组成一个数据仓库，所以对单个数据集市的设计必须采用分布式和增量式设计。其他事情都意味着某种折中，所以必须在某些方面有所改进。

- **它是能保护信息资产安全的安全堡垒**

数据仓库不仅能有效地控制数据的存取，而且能为其所有者提供非常大的可见度，使后者能够了解数据的使用和误用情况，即使在它已离开数据仓库以后也能够实现这一点。其他事情都意味着某种折中，所以必须在某些方面有所改进。

- **它是决策的基础**

数据仓库拥有用于支持决策活动的合适数据。从数据仓库中只有一种真实的输出（即用于决策）。在数据仓库提供了相关证据以后，就可以做出决策。数据仓库的最初标签是“决策支持系统”，它仍然最适合用来描述我们正在试图创建的东西。

本书的写作目的

当本书继续获得成功以后，大型数据仓库的设计人员和管理人员就能够更快地实现其目标。他们将会创建各种高效的数据仓库，这些数据仓库的目标与本书前面章节中所概述的数据仓库目标能够很好地匹配，同时在该过程中所犯的错误也会更少。幸好不必重新回头，并且发现“先前所拥有的”各种真理。

本书试图尽可能多地从技术角度去探讨数据仓库这样一个大主题，而不被面向特定产品销售商的具体细节所纠缠。对于从事数据仓库市场营销工作的人员来说，他们的一个兴趣点肯定是在理解所有数据仓库职责时所必需的知识宽度上。我们非常强烈地感觉到在这方面有必要保持较宽泛的观点，主要是因为数据仓库具有不断进化的特征。即使数据仓库已经超越了文本和数字数据这些基础概念，或者依靠关系型数据库技术，本书中所提及的大多数原则仍然适用，因为数据仓库项目组的使命从字面意义上看，最重要的就是要创建一个决策支持系统。

在拥有适量的结构和规范时，就可以为创建复杂的大型数据仓库提供很大帮助。因此，我们打算通过本书介绍这些结构和规范，希望读者能够理解和参与整个业务维生命周期法，同时还将这种观点灌输给整个组织。数据仓库在许多方面体现了信息系统中的一个重要思想，即收集组织信息，并使之