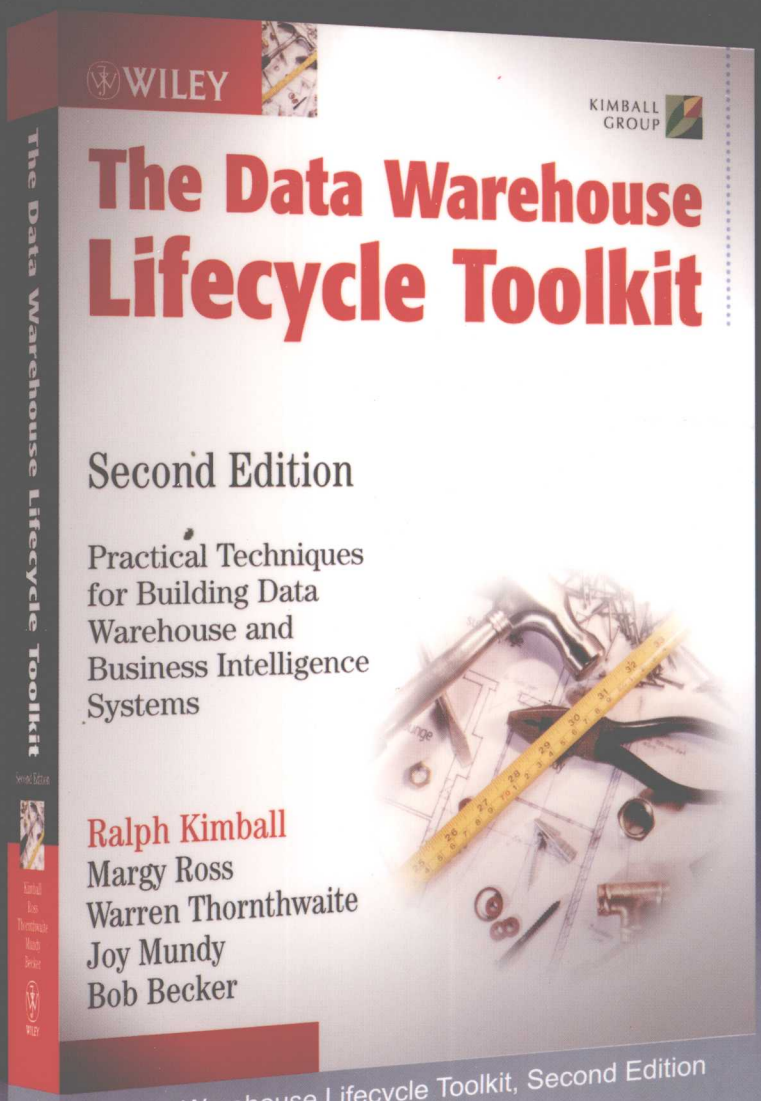
国外计算机科学经典教材

数据仓库生命周期工具箱（第二版）

(美) Ralph Kimball 等著　　唐富年 孙媛媛 译



WILEY
清华大学出版社

国外计算机科学经典教材

数据仓库生命周期工具箱

(第二版)

(美) Ralph Kimball 等著

唐富年 孙媛媛 译

清华大学出版社

北 京

Ralph Kimball, Margy Ross, Warren Thornthwaite, Joy Mundy, Bob Becker
The Data Warehouse Lifecycle Toolkit, Second Edition
EISBN: 978-0-470-14977-5

北京市版权局著作权合同登记号　图字：01-2009-4344

图书在版编目(CIP)数据

数据仓库生命周期工具箱(第二版)/ (美) 金博尔(Kimball，R.)等著；唐富年，孙媛媛 译.
—北京：清华大学出版社，2009.9
书名原文：The Data Warehouse Lifecycle Toolkit，Second Edition
ISBN 978-7-302-20374-2
I. 数…　II. ①金…　②唐…　③孙…　III. 数据库系统　IV. TP311.13
中国版本图书馆 CIP 数据核字(2009)第 100738 号

责任编辑：王　军　李楷平
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：清华大学印刷厂
装 订 者：三河市溧源装订厂
经　　销：全国新华书店
开　　本：185×260　**印　张**：31　**字　数**：754 千字
版　　次：2009 年 9 月第 1 版　　**印　　次**：2009 年 9 月第1 次印刷
印　　数：1～4000
定　　价：68.00 元

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

关于作者

本书几位作者的职业生涯都有着非常相似的轨迹。每个作者在数据仓库和商业智能(DW/BI)领域从事咨询和培训的时间都超过了 15 年。大多数作者都曾经在 Metaphor 计算机系统公司(20 世纪 80 年代一家具有开创性的提供决策支持系统产品的厂商)中共事过。所有的作者都是 Kimball 集团的成员，都在 Kimball 大学讲授课程。他们经常向 *Intelligent Enterprise* 杂志和其他的业界刊物供稿，而且大多数作者以前都参与过工具箱系列丛书的撰写。

Ralph Kimball 创建了 Kimball 集团。从 20 世纪 80 年代中期开始，他就是 DW/BI 行业维度方法的领导者，并且培养了 10 000 多位 IT 专业人士。Ralph 在斯坦福大学(Stanford University)获得过电子工程专业博士学位。

Margy Ross 是 Kimball 集团的总裁。她从 1982 年就开始专门从事 DW/BI 方面的研究，特别强调业务需求分析和维度建模。Margy 在西北大学(Northwestern University)获得过工业管理学学士学位。

Warren Thornthwaite 从 1980 年开始从事 DW/BI 方面的研究。他刚开始负责管理 Metaphor 公司的咨询机构，后来为斯坦福大学和 WebTV 工作。Warren 在密西根大学(University of Michigan)获得过传媒研究学士学位，并且从宾夕法尼亚大学沃顿商学院(University of Pennsylvania's Wharton School)获得了 MBA 学位。

Joy Mundy 从 1992 年开始从事 DW/BI 方面的研究，在斯坦福、Web TV 和微软公司的 SQL Server 产品开发机构中都工作过一段时间。Joy 在塔夫斯大学(Tufts University)获得过经济学学士学位，并且在斯坦福大学获得了工程经济系统硕士学位。

Bob Becker 从 1989 年就开始帮助多个行业的客户解决 DW/BI 方面的难题，并提供相应的解决方案，还为卫生保健机构做了大量的工作。Bob 从明尼苏达州商学院(Minnesota's School of Business)获得了市场营销学商学士学位。

译 者 序

本书是数据仓库和商业智能领域的又一部经典著作，作者 Kimball 等人在数据仓库领域享有很高的声誉。本书的作者都长期工作在数据仓库/商业智能系统开发的第一线，他们将自己多年的经验和感悟写入到了本书的字里行间。全书讲述了整个 Kimball 生命周期过程各个环节的具体工作，从业务需求的视角，引导读者全面认识数据仓库/商业智能系统的开发。

本书并不是一本为刚刚涉足数据仓库领域的新手准备的入门教材，我们建议您在阅读本书之前最好能有一些数据仓库或数据库领域的基本知识，这对把握本书的内容有很多好处。在书中，作者使用了一种通俗而明晰的手法进行内容的讲述，因此译者在翻译时也尽量使用比较通俗的语言。本书的写作风格匠心独运，在内容安排上也体现出了 Kimball 生命周期中的“迭代”思想。全书在章节上有总有分，一层层展开，一步步深入。一些读者刚刚拿到这本书时也许难以习惯它的内容布局，但是正如作者所说，只要坚持读完全书，最终必然会理解数据仓库/商业智能系统开发工作的总体流程和关键环节。

本书在每一章开始都说明了这一章的内容适合于哪些读者，在每一章的结尾都进行了详细而具有针对性的总结。读者可以通读全书，也可以根据自身的需要选读其中的部分章节。目前市场上有关数据仓库的书籍非常多，而且各自的侧重点都有所不同，恐怕难以评论孰优孰劣。但是本书的出现，也许会让您有耳目一新的感觉，因为本书的出发点是站在业务的角度，而并没有单纯从技术的角度来安排各个章节的内容，这就使书中的内容更贴近实际，也更具有可操作性。

在本书的翻译过程中，我们参考了很多数据仓库和商业智能方面的书籍，这对我们准确把握本书的内容起到了重要作用。遗憾的是，在数据仓库和商业智能领域至今仍然有大量的术语没有统一的标准译法，在翻译过程中我们对这些术语尽量都使用了比较常用的译法。在此，非常感谢本书第一版的译者们，他们的工作具有很高的借鉴价值。

本书由唐富年、孙媛媛翻译。同时，也非常感谢国防科技大学信息系统与管理学院的张文煜博士，他为本书的翻译提出了一些很宝贵的意见。肖国尊负责翻译质量和进度的控制，以及翻译思想的指导，在此予以衷心感谢。鉴于译者水平有限，译文中难免存在错漏之处，还望广大读者谅解并不吝指正。敬请读者将反馈意见发至邮箱 be-flying@sohu.com。

前　言

在《数据仓库生命周期工具箱》第一版出版之后的九年里，数据仓库产业已经发生了显著的变化。现在，数据仓库产业已经变得十分成熟，并且得到了商业界的接受和认可。在这九年里，硬件和软件的发展取得了令人难以置信的成就。我们已经开始谈论“TB”字节而不再是“GB”字节。但是，数据仓库的任务基本上没有什么改变。

许多人所在的机构里有数千位的数据仓库用户，从业务决策者到一般的数据仓库用户，再到市场营销和财务用户中的骨干成员。事实上，操作方面的迫切需求是数据仓库研究中的最热点问题，而且每个人都坚持认为他们需要“实时”数据。在数据仓库变得越来越重要、越来越直观的同时，用户反复提出保密性、安全性和合规性方面的需求。业务用户正在逐步意识到高质量数据的价值，这和传统制造业注重质量管理是一样的道理。最后，可能也是最重要的，我们为自己所从事的这个行业起了一个新的名称，这个名称反映了我们的真正目的。它就是：商业智能(Business Intelligence)。为了强调这一点，在本书的大多数地方，我们都把您要创建的整个系统叫做 DW/BI 系统。

商业智能的这种转变将主动权移交到了业务用户手中，而不再是由 IT 人员掌握主动权。但同时这个转变将全部注意力都集中到了数据仓库的使命上：它是商业智能必需的平台。数据仓库需要做繁重的工作，它从源系统中取得数据，对数据进行清洗，并将数据组织起来使普通的业务用户能够看懂它。当然，我们力争实现世界级的商业智能，但是世界级的商业智能需要您拥有一个世界级的数据仓库。反过来说，一个数据仓库没有商业智能将会遭遇彻底的失败。

本书是 DW/BI 系统的设计人员、管理人员和所有者在实际工作中的指南。我们尽可能使本书的内容非常具体和实用，以便将这本书与其他 DW/BI 书籍区分开来。被本书的内容搞得眼花缭乱并没有关系，我们希望您一直坚持读下去，最终必然会达到预定目标。这本书描述了一个条理清晰的 DW/BI 系统框架，这个框架从确定整个企业 DW/BI 系统的初始范围开始，经过详细的开发和部署步骤，一直到最后计划下一个阶段的工作。

全世界安装有好几万个功能各异的数据仓库。很多 DW/BI 系统的所有者都是完全按照生命周期的思想来进行开发的。或许从生命周期思想中可以得到的最大收获就是：每个 DW/BI 系统都是不断发展变化的，它永远都不是静止的，也决不会停止转变。新的商业需

求会不断涌现，新的管理者和执行官也会对 DW/BI 系统提出一些不可预知的要求，还会有新的可用的数据源加入到系统中。至少，DW/BI 系统需要根据所处机构的变化而同步调地变化。稳定的机构也会要求 DW/BI 系统取得适度的演化，不断变化的动态机构则会使 DW/BI 系统的任务变得更具有挑战性。

考虑到 DW/BI 系统具有不断演化的特性，我们需要灵活可变的、适应性强的设计方法，还必须同时扮演 DBA 和 MBA 的角色。我们需要将来自单个业务过程的小块数据连成大块数据，从而形成企业级的数据仓库。同时，还要求对 DW/BI 系统所做的改变始终是适度的。一个适度的改变不会使以前的数据或先前的应用程序无效。

本书结构

本书有两大基本主题。第一个主题就是 Kimball 生命周期方法。您可能会问："是什么使得 Kimball 生命周期方法与其他方法不同？"最简单的答案是我们从业务用户的角度开始，找出他们完成工作需要什么，并以此来建立 DW/BI 系统。有了这些需求之后，我们逐步向下使报表、应用程序、数据库和软件系统地进行工作，最后再深入到底层设施的物理层。这与技术驱动的方法形成鲜明的对比——其顺序正好相反。在 20 世纪 90 年代初期，一些 IT 工作室并不知道怎样使用我们这种面向业务和用户的方法。但是随着 2008 年本书的出版，"商业智能"这个名称本身就说明了这一切，即应该由用户和业务来驱动数据仓库。

第二个主题是"总线架构"。本书中介绍了如何进行单个业务过程的连续迭代，使读者最终能够创建一个企业 DW/BI 系统。在本书中，您会看到我们将维度模型作为一种向业务用户表现数据的可靠方式。推荐这种方法只有一个原因：它确实能够满足业务用户的愿望，简单并且具有高效的查询性能。我们由衷地感谢您能够选择本书所讲述的维度建模方法。最后，您可以使用任何您认为比较合适的方式将数据呈现给用户。这不应当由我们来决定，而是应当由用户来决定。

本书中涵盖了上述观点，提供了能够辅助用户完成工作的各种具有实际价值的技能和实用的工具。通过这种方式，希望能将我们自 1982 年以来在建立 DW/BI 系统的过程中积累的想法和经验都讲授给您。

本书读者对象

本书的主要读者应该是那些真正需要在实现"作为商业智能应用平台的数据仓库"的过程中负责创建和管理工作的设计人员或者管理人员。因为"作为商业智能应用平台的数据仓库"这句话十分冗长而拗口，因此在提到整个系统时我们都使用"DW/BI"这一名称，它说明您需要负责从初始源系统获得数据一直到将数据显示到业务用户屏幕的整个过程。

尽管本书包含一些介绍性的内容，但是我们认为这本书对于已经对数据仓库技术有一定接触的 IT 专业人员来说将非常有用。在 2002 年出版的另一本相关书籍，是由 Ralph Kimball 和 Margy Ross 编写的 *The Data Warehouse Toolkit [Second Edition]*，该书更加深入而具体地讲述了维度建模。

通过设计和交付一个真正的数据仓库，您可能已经积累了一些经验并且形成了自己的观点，这就是最好的知识背景！开发一个实际的数据仓库所积累的经验是任何其他方式都无法替代的。在将自己的“杰作”提交给一群要求苛刻的业务用户时，我们这些作者都曾有过一些令人感到羞愧的经历。令人难以接受的是，大多数用户的实际工作与技术毫不相干，他们甚至可能特别不喜欢技术。但是如果我们的技术易于使用，并且能够为用户提供明显的使用价值，那么业务用户还是会使用我们的技术的。

本书要求有一定的专业知识。其中，有关设计技巧和体系结构方面的论述毋庸置疑地会引入一些您未曾遇到过的专业术语。我们已经对本书进行了精心梳理，以确保大多数技术方面的主题都是读者应该能够理解的。我们尽量使本书不会因为内容本身的原因而使您陷入细节上的困扰。在本书后面的 DW/BI 的术语表中，将简要地解释我们在书中所使用的术语。

尽管我们希望读者能够完整地阅读本书来理解 Kimball 生命周期的全过程，但是我们在每一章的开始也会强调该章所主要针对的读者，这样您就可以更好地判断哪些内容需要精读，哪些内容可以跳过。希望您的经验和看法使您搭建起自己的框架，这样我们的观点就可以被串联在一起了。阅读完第一章以后，您将会看到在建立一个 DW/BI 系统时必须按照三条并行的路线推进：技术、数据和商业智能应用程序。我们还在每一章开头的“You Are Here”图中都指明了这三条路线。尽管这三条路线之间明显会互相影响，但是它们的开发应该以并行方式和异步方式进行。

由于图书的内容必然是按照线性方式进行编排的，所以我们不得不线性地介绍 Kimball 生命周期中的所有步骤，就像这些步骤是以某种固定的次序发生一样。希望在读完本书以后，您能够想象出这些步骤在现实世界中是具有更现实、更复杂的关系的。读完本书以后，当您的项目进行到某个特定阶段时，请再返回到相应的章节并重新仔细地阅读其内容。这也就是为什么我们将其称作生命周期工具箱的原因。

这本书与第一版有什么不同

与第一版相比，第二版生命周期工具箱的内容有了明显的更新和重组。前三章可以帮助您理解整个 Kimball 生命周期过程，并确保您的工作已经满足继续向前推进所必需的条件。然后，我们努力使有关复杂架构的讨论更具有实际价值，并且更加紧密地将架构和 Kimball 生命周期中各项活动的次序联系起来。在第 4 章我们细致地讲述了 DW/BI 系统的完整架构，包括从原始数据的提取到最后将数据显示到业务用户屏幕上的整个过程。在第 5 章我们讲述了怎样为这个技术架构创建一个详细的计划和如何进行产品选择。然后，从第 6 章到第 12 章，我们沿着三条主线(数据库设计、ETL 系统和 BI 应用程序)系统地展开，先从概念上进行介绍，随后又从物理上进行了介绍。在最后的两章里，讲述了如何将这个精心设计的庞然大物部署到实际业务环境中，并且讨论了在第一轮实现之后怎样扩展 DW/BI 系统。

希望我们对数据仓库和商业智能的热情能够贯穿本书的始终。DW/BI 所面临的挑战令人着迷，而且也值得探索。毋庸置疑的是，数年之后当 DW/BI 厂商对他们的产品加以更新时，所有以前的东西都会被取代，相应的名称也会改变。但是我们的任务仍然不会改变，那就是：为业务用户提供数据和分析结果，使他们能够更好地进行业务决策。

目　录