国外计算机科学经典教材

The Microsoft Data Warehouse Toolkit: With SQL Server 2005 and the Microsoft Business Intelligence Toolset

数据仓库工具箱

——面向 SQL Server 2005 和 Microsoft 商业智能工具集

Joy Mundy
(美) Warren Thornthwaite 著
Ralph Kimball
闫雷鸣 冯 飞 译
杨大川 审校

WILEY
清华大学出版社

国外计算机科学经典教材

数据仓库工具箱

——面向 SQL Server 2005 和 Microsoft 商业智能工具集

Joy Mundy
(美) Warren Thornthwaite 著
Ralph Kimball
闫雷鸣 冯飞 译
杨大川 审校

清华大学出版社

Joy Mundy, Warren Thornthwaite, Ralph Kimball
The Microsoft Data Warehouse Toolkit: With SQL Server 2005 and the Microsoft Business Intelligence Toolset
EISBN: 0-471-26715-5

图书在版编目(CIP)数据

数据仓库工具箱——面向 SQL Server 2005 和 Microsoft 商业智能工具集/(美)曼蒂(Mundy, J.), (美)桑斯维特(Thornthwaite, W.), (美)金伯尔(Kimball, R.)著；闫雷鸣，冯飞 译. —北京：清华大学出版社，2007.12
(国外计算机科学经典教材)
书名原文：The Microsoft Data Warehouse Toolkit: With SQL Server 2005 and the Microsoft Business Intelligence Toolset
ISBN 978-7-302-16379-4
Ⅰ. 数… Ⅱ. ①曼… ②桑… ③金… ④闫… ⑤冯… Ⅲ. 数据库系统 Ⅳ. TP311.13
中国版本图书馆 CIP 数据核字(2007)第 168762 号

责任编辑：王 军 于 平
装帧设计：孔祥丰
责任校对：成凤进
责任印制：李红英
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
c—service@tup.tsinghua.edu.cn
社 总 机：010-62770175 **邮购热线**：010-62786544
投稿咨询：010-62772015 **客户服务**：010-62776969
印 刷 者：北京市世界知识印刷厂
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**：37.5 **字 数**：913 千字
版 次：2007 年 12 月第 1 版 **印 次**：2007 年 12 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、Thomson 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、Thomson 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

Joy Mundy 是 Kimball Group 的成员之一，自 1992 年起就一直从事商业智能系统和技术的开发、咨询、演讲和写作的工作。Joy 最初作为银行和金融领域的分析师开始其职业生涯，并成为了商业智能领域的超级用户。在 1992 年，她加入了斯坦福大学的数据仓库小组，接下来，她和别人一起创建了一个数据仓库咨询公司——InfoDynamics LLC；然后她加入了 Microsoft WebTV，开发闭环分析应用程序以及一个被打包的商业智能系统。

从 2000 年到 2004 年，Joy 一直在 Microsoft SQL Server Business Intelligence 产品研发小组工作，她管理着一个能够开发用于在 Microsoft 平台上构建商业智能系统的最佳实践的小组。Joy 在塔夫茨大学获得经济学学士学位，然后在斯坦福大学获得工程经济系统硕士学位。

Warren Thornthwaite 是 Kimball Group 的成员之一，自 1980 年起就一直从事共建决策支持和数据仓库系统的工作。Warren 和别人合著了畅销书 *Data Warehouse Lifecycle Toolkit* (Wiley, 1998)。

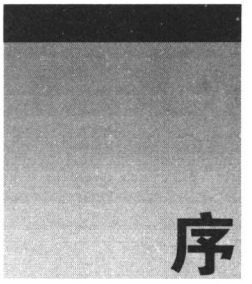
Warren 从 1983 年起在 Metaphor Computer Systems 工作了 8 年，在那里，他管理一个咨询公司并实现了许多主要的数据仓库系统。离开 Metaphor 后，Warren 管理斯坦福大学企业级的数据仓库系统开发，接着他和别人合作创立了一个数据仓库咨询公司——InfoDynamics LLC。在回到咨询领域之前，他又加入了 WebTV 小组帮助构建了数以 TB 的面向客户的世界级数据仓库。

除了为不同工业领域设计数据仓库外，Warren 在帮助客户开发可扩展的、实用的信息存取体系结构方面也具有广泛的经验。Warren 从宾夕法尼亚州的 Wharton School 大学获得 Decision Sciences 的 MBA 学位，从 Michigan 大学获得 Communications Studies 的学士学位。

Ralph Kimball 博士是 Kimball Group 的创立者，自 1972 年起就一直从事设计信息系统和数据仓库的工作。

Ralph 毕业于斯坦福大学的 Electrical Engineering 系，他的博士论文是关于设计辅导数学学生的人机系统。在 1972 年，他作为研究科学家加入了 Xerox Palo Alto Research Center。在 Xerox 的接下来的十年中，他成为了 Xerox Star 工作站的开发经理以及产品销售经理，这个工作站是使用视窗、图标和鼠标的最早的商业产品。因为在 Xerox 的这一工作，他获得了 IEEE Human Factors Society 用户接口设计的 Alexander Williams 奖。

接下来的几年，Ralph 是 Xerox 的副总裁并且是 Metaphor Computer Systems 的创始小组成员。在 1982~1986 年，Metaphor 安装了许多客户-服务器数据仓库系统。在 1986 年，Ralph 创立了 Red Brick Systems，他开发了用于决策支持的第一个高性能的关系型数据库。自 1993 年起，Ralph 就开始设计数据仓库系统，编写最畅销的数据仓库书籍并向一万多个 IT 专业人员讲授数据仓库技术。



序

作为工程师，作为 Kimball Group 中的成员，我们希望构建那些感兴趣的过程，包括构建一个数据仓库的过程，以及最终用户如何从数据仓库和生成的 BI 系统中获取值的过程。同时，我们惊叹于设计任务的艰难。在过去 20 年中，在一直构建数据仓库的同时，我们也在寻找简化和划分设计过程的方法。当我们看到相同的设计过程重复地出现时，我们会给予这种技术一个名称，然后尝试以一种清晰的方式来解释它。这些设计技术的集合称为 Kimball Method，本书详细地介绍了这些技术。

Kimball Method 的发展积聚了重要的因素，在过去 10 年中，我们培训了 10 000 多名数据仓库设计人员，并且卖出了 200 000 多本书，这些书都是阐述 Kimball Method 的。最近，我们看到 Kimball Method 已经被数据仓库领域的主要技术供应商所采纳。当您阅读本书时，要时刻关注 Microsof 公司标有“the Slowly Changing Dimension wizard”(渐变维度向导)字样的 SQL Server 2005 家族的产品特性。我们采用的维度方法已经成为业界工具集中主导的主题。

即使采用了 Kimball Method，构建数据仓库及其依赖的 BI 系统仍然需要对前景和判断进行深入的训练。尽管这本书可能和十几本关于 Microsoft SQL Server 2005 的书差不多厚，但却和它们完全不同。这是一本真实的“判断”书籍，而不是一本“如何做”的书籍。在写这本书时，Joy 和 Warren 作为数据仓库设计师和帮助推出 SQL Server 2005 的前 Microsoft 雇员，应用了他们独特的观点。

我希望您能够赞赏 Joy 和 Warren 所采纳的经过深思熟虑的方法，当一个主题过于复杂时，没有深究。在有些地方，当 Microsoft 可以使一个特征更简单时，他们会用更多的语言来对此进行描述。本书主要尝试将数据仓库设计人员在使用 SQL Server 2005 的每一阶段应该考虑的问题进行了可视化。我相信 Joy 和 Warren 已经成功地将高层设计判断和关于这些工具细节的许多有用的注释结合在了一起。我很荣幸自己在本书介绍 Kimball Method 的部分时起了一定的作用。

Ralph Kimball

审校者序

和大家讨论几个问题，希望通过以下简单的几段话，读者能够对本书的价值有所了解。

- 什么是商业智能？

商业智能(Business Intelligence)是一种解决方案，它的目的是把用户积累下来的、大量的数据转化为业务人员容易理解的信息，进而辅助决策。

数据库中存在的是数据，对于业务人员来说，只是一些无法看懂的天书，没有人会去拿放大镜分析数据库服务器硬盘上的磁轨。他们需要的是信息。那么，我们以前是如何解决这个矛盾的呢？一般的答案是报表系统。简单说，业务人员看到的是美观的界面、便捷的操作。鼠标点击后，报表系统生成 SQL 语句，数据库服务器收到以后，返回所需要的信息。不错，报表系统已经可以称作是 BI 了，它是 BI 的低端实现。

现在国外的企业，大部分已经进入了更深一些层次的商业智能，叫做数据分析，即基于多维数据库的在线分析系统(OLAP)。还有一些企业已经开始进入更深一个层次的商业智能，叫做数据挖掘(Data Mining)。

- 什么是数据仓库？

通常，数据仓库有两层含义。

一个完整的商业智能项目，一般来说，要经过以下步骤：通过 ETL 把各类数据导入一个中央存储区域(数据仓库)，然后建立多维模型，在此基础之上，进行报表、数据分析和数据挖掘不同层次的应用。

狭义来说，上述的“中央存储区域”叫做数据仓库。

广义来说，整个商业智能的各个环节，都可以称为数据仓库。

值得一提的是，本书讨论的内容包括了从 ETL、数据仓库模型一直到多维模型、数据挖掘等，因此这里的数据仓库是广义的定义。如果您愿意把这本书称作“商业智能工具箱”，也是完全可以的。

- 一个商业智能是否成功，关键在于什么？

如果您去问一位资深的程序员：“我的程序用 Java 语言开发能成功还是 C#语言开发能成功？”，答案很可能是：“语言本身并不重要，重要的是如何设计、如何搭建合理的程序架构……”

同样的，如果您去问一位商业智能专家：“我的商业智能项目应该采购什么产品？”他会告诉您，产品不是最重要的，商业智能工具的使用方法也不是最重要的，成功的关键

在于如何分析需求，如何设计模型，如何在方法论的层面解决问题。

这本书与绝大多数其他技术书籍的不同点在于，这本书试图阐述商业智能方法论层面的知识，或者说“该做什么？”，而不是“如何做？”。

我在国外学习、工作了多年之后，深感商业智能即将成为未来几年 IT 领域的核心价值，因此从 2003 开始创建了北京迈思奇科技有限公司，致力于将国外的先进商业智能技术和工具引进国内，帮助国内的企业提高数据分析效率、增强竞争实力。公司成立四年来，与微软密切合作，通过近百次讲座和培训，为企业培养商业智能专业人员；同时，在承担商业智能项目开发的过程中，公司也积累了优秀的团队和丰富的项目案例，创立了国内一流的品牌。

感谢清华大学出版社，及时为我们引进了这部优秀的教材；感谢本书的译者，准确而清晰地传达了原著的精华；也感谢迈思奇公司参与本书审校工作的各位 BI 咨询顾问。

北京迈思奇科技有限公司　杨大川

前 言

本书描述了如何使用 Microsoft SQL Server 2005 产品集构建一个成功的商业智能系统以及数据仓库数据库，这里的关键字是“成功的”。

0.1 数据仓库和商业智能系统

数据仓库和商业智能的作用在于，为业务人员提供制定操作性和战略性业务决策所需的信息和工具。我们将详细剖析这方面的问题，以便您能真实了解将要采纳的决策的性质和规模。

首先，客户一般是指公司的业务人员。但是对您来说，并非所有的业务用户都具有同等的重要性——您对于那些制定战略性业务决策的人可能更感兴趣。为什么？因为这就是真正产生利润的地方。一个非常好的业务决定可能意味着许多公司的数百万美元。您主要的客户是主管人员、经理以及公司的分析师，因此，数据仓库和商业智能(DW/BI)系统影响深远、意义重大。

战略性也意味着重要性。这些是可以决定公司成败的决策。因此，DW/BI 系统是一个高风险的努力。当制定了某个重要决定时，有些人将会成功而有些人将会失败，因此 DW/BI 系统也是高度战略性的努力。

DW/BI 系统正日益支持着操作性决策，特别是在决策制定者需要从多个数据源中查询历史数据或集成数据的地方。许多“分析型应用程序”都支持这一操作。从技术角度看，不管制定的决策是战略性的还是操作性的，您都需要提供必要的信息来制定这些决策。任何给定的决策都需要一个独特的信息子集。您需要跨公司并且可能从公司外部提取数据来构建一个信息基础结构，然后清理、排列和重构这些数据来使得它们更灵活和更可用。尽管大部分事务系统模块和某种类型的数据，例如收到的账单、订单或账目一起工作，DW/BI 系统最终必须将它们集成在一起。因此，DW/BI 系统要求技术上很复杂的数据收集和管理。

最后需要给业务决策制定者提供使用这些数据的工具。在这样的情形下，工具并不仅仅意味着软件，这意味着业务用户需要了解的所有事情，哪些信息是可用的，查找需要的子集以及将数据结构化来阐明潜在的业务动态。因此，工具意味着培训、文档、支持，以及即席查询查询工具、报告和分析型应用程序。

让我们一起回顾 DW/BI 系统：

- 是高配置的和高影响的；
- 是高风险的；
- 是高度战略性的；
- 要求技术上很复杂的数据收集和管理；
- 要求强化的用户访问、培训和支持。

创建和管理 DW/BI 系统是一项具有挑战性的任务，我们希望以您所了解的全部知识来接受这一任务。以我们的经验来看，如果您事先被预警，将很容易应对这些挑战。

我们并不想使您气馁，而是要在您跳进深水之前警告您，以我们的经验来看，所有使得数据仓库具有挑战性的原因也就是使得它成为一个很有趣且令人兴奋的项目。

0.1.1 Kimball Group

尽管构建和管理一个成功的 DW/BI 系统是具有挑战性的，但也有一些增加成功可能性的方法，那就是使用 Kimball Group 所提供的方法。我们已经在 DW/BI 领域工作了 20 多年，本书的作者也是 Kimball Group 的成员，整个职业生涯作为销售商、顾问、实现者和用户，一直都从事着数据仓库和商业智能系统的工作。我们的格言是“Practical techniques—proven results(实践的技术证明了结果)”，我们共同的动力是指明构建和管理一个成功的 DW/BI 系统的最好方式。我们也是热心的老师，衷心希望您通过学习本书获得成功并且避免我们和其他人犯过的错误。

0.1.2 本书目标

数据仓库和商业智能至少自 1970 年就具有相似的形式，并且持续享受着无限的技术生命周期。在 1995 年，我们的主要作者构建了第一个顾问公司，其中的作者之一认为数据仓库已经结束了，这个浪潮已经开始回落。幸运的是，我们在找到工作之前获得了更多的项目。12 年后，数据仓库和商业智能依然很强大，事实上，仅仅在过去几年我们才看到它们在工业上的成熟。

成熟市场的一个标志就是单源提供者的出现——对不愿冒风险的公司来说这是一种安全的选择。数据仓库技术涵盖了从深奥源系统知识到用户接口设计以及具有最好实践的 BI 应用。尽管许多销售商在最近几年都争着把自己放在端到端的提供者位置上，但对于我们来说，很显然，数据仓库销售商确实是那些可以提供端到端解决方案的人。在 2001 年，当我们首次讨论这本书时，我们已经感觉到 Microsoft 要以一个诱人的价格强行将一个可行的、单源数据仓库系统提供者的概念加入到现实世界中。

我们相信向单源提供者的转变意味着必须将 Kimball Method 技术扩展到特定的产品级，使其可以直接投放单源提供者市场。我们选择 Microsoft 工具集作为测试样例有两个原因，首先，SQL Server 2005 是一个强大的 BI 平台，Microsoft 自 20 世纪 90 年代中期投资 Analysis Services 引擎以来，就一直在扩展和增强商业智能方面投资巨大。投资的级别也因此巨大地翻升。随着 SQL Server 2005 开发的开始，SQL Server 2005 开发团队增长到 200 人，Microsoft 对于将商业智能引入主流市场很认真。其次，两位作者都从 1997 到 2002 或 2004 在 Microsoft 工作，特别地，Joy 曾是 SQL Server Business Intelligence 开发团队中 SQL

Server BI Best Practices 组的经理，这可以给予我们一系列很强的工作关系以及访问关键的支持资源。

0.1.3 本书读者对象

本书覆盖了整个数据仓库生命周期，因而可以给数据仓库团队的每个成员提供有用的指导，从项目经理到业务分析师、数据建模者、ETL 开发者、DBA，分析型应用开发人员甚至业务用户都可以从本书中受益。我们相信本书对从事 Microsoft SQL Server 2005 数据仓库项目的任何人都非常有价值。

本书的主要读者是在 Microsoft SQL Server 平台上启动项目的新的 DW/BI 团队，我们假定您并没有构建 DW/BI 系统的经验，但假定您对 Microsoft 世界有一个基本的认识：操作系统、基础设施组件以及资源。我们也假定您对关系数据库(表、列和简单 SQL)有一个基本认识，并且对 SQL Server 2000 关系数据库有一定认识，尽管这并不是一个必备条件。贯穿全书，我们提供了许多其他书和资源的参考。

第二个读者群是有 Kimball Method DW/BI 使用经验但首次接触 Microsoft SQL Server 2005 工具集的读者，这些读者可能需要阅读一些资料以便了解基础结构，特别是如果您从来没有使用过 Windows Server 更需如此。我们将指出对于那些曾经阅读过我们的 Toolkit 书籍以及实践过我们的方法的读者需要复习哪些部分和章节，不过再次阅读这些材料并没有坏处。

不管您的背景如何，如果您从一个新项目开始将从本书受益匪浅。尽管我们确实提供了运转现有的数据仓库的建议，但在理想情况下，您不会对任何已有的数据仓库或数据集市满意，至少在新系统部署后对仍然留在原处的系统不会满意。

0.2 业务维度生命周期

在深入一个项目后，我们都会感到内心恐慌，因为我们面前努力的范围和规模将会比我们在外围的想象花费更多的精力。许多 BI/DW 系统都从这样的概念开始：移动一些数据到新机器中，清理数据然后开发一些报告。这听起来很不错，但经过六周后或最多两个月的努力，在您意识到需要建立一座桥之前您已经身陷河中。

避免这种恐慌的最好方式是在您跳入水中之前告诉您应该去哪里，通过提供一些路标和方向可以将您安全地领过不熟悉的区域，这些路标和方向将告诉您哪些地方是安全的以及前途中会有哪些危险区域。本书就是 Microsoft SQL Server BI/DW 系统项目的路标。它遵循了 *The Data Warehouse Lifecycle Toolkit*(Wiley，1998)首先描述的业务维度生命周期中的基本流程。本书汇集了根据我们的经验，很好地描述了生命周期的这些步骤、任务和依赖关系。生命周期是一个基于四个主要原则的迭代的方法：

- 专注于业务：集中识别业务需求和它们的关联。努力开发牢靠的业务关系并且让您的业务感和咨询技能更敏锐；
- 构建一个信息基础设施：设计一个单一的、集成的容易使用的高性能的信息基础设施，旨在满足您在跨企业中识别的、更加广泛的业务需求；

- 进行有意义的增量发布：以增量方式构建数据仓库，可以以 6~12 个月的时间来发布。使用可以清晰识别的业务值来决定增量的执行次序；
- 发布整个解决方案：提供所有必要的元素来传递值给业务用户。这意味着一个稳固的、设计很好的、高质量的可用数据仓库正在开始中。您也必须发布即席查询工具、报告的应用以及高级的分析、训练、支持、Web 站点和文档。

本书通过使用业务维度生命周期帮助您遵循构建 DW/BI 系统的四个原则，这四个原则已经加入到生命周期中。理解业务维度生命周期的秘密就清晰地隐藏于它的名字中：它是基于业务的，采用维度方法用于设计展现给最终用户的数据模型，而且它是一个真实的生命周期。

0.2.1 生命周期跟踪和任务区域

BI/DW 系统是一个复杂的实体，并且构建这种系统的方法必须能够简化此复杂性。图 0-1 展现了生命周期。13 个方框显示了关于构建一个成功的数据仓库和这些任务之间的主要依赖性的主要任务区域。

说明：迟做总比什么都不做要好

发布整个解决方案是我们的根本原则之一，我们甚至不会考虑在没有发布现在所谓的商业智能层之前构建一个数据仓库。在经历过仅仅专注于创建数据仓库数据库的失败项目后，在 1990 年后期其他工业才逐步认识到这一点。

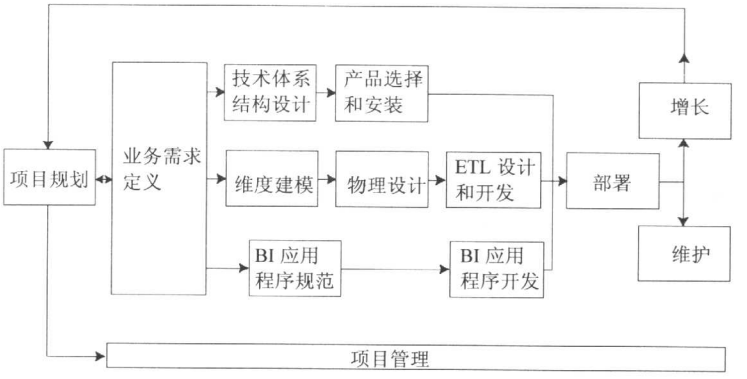


图 0-1 业务维度生命周期

在生命周期这一级可以进行多方观察，首先，注意到业务维度生命周期的业务需求定义方框的中心角色。业务需求提供了紧随其后的三个轨迹的基础，它们也影响着项目计划，因此箭头指向了后面的项目规划方框。经常会以修改一个基于业务需求和优先权的更详细的理解结束。其次，生命周期中间的三个轨迹集中了三个不同的领域。

- 顶部的轨迹是关于技术的。这些任务主要是关于计划使用哪部分 Microsoft 技术以及如何安装和配置它们。

- 中间的轨迹是关于数据的。在数据轨迹中，将设计和实例化维度模型，然后开发ETL(提取、转换和加载)系统来填充它。可以将数据轨迹视作“构建数据仓库数据库”，尽管数据仓库直到将生命周期任务的其余部分包围上时才会成功。
- 底部的轨迹是关于商业智能应用的。在这些任务中将为业务用户设计和开发BI应用程序。

当开始部署系统时这些轨迹将合并，这是一个非常微妙的时刻，因为仅有一次机会可以获得第一好印象。维护DW/BI系统并不在部署之后开始，需要有一定的能力来设计系统并要有工具来维护它。项目增长阶段链接到的箭头有一些主要的隐含之意，生命周期的增量方法是发布业务值的基本元素。见图0-1。

在整个生命周期的下面是项目管理方框，这里要记住的最重要的事情就是您需要一个领导者，并且他能够访问高级管理。团队领导者是那些很难找到的理想的人选之一，他们是可以同技术人员和业务人员，包括公司最高级的执行者进行有效沟通的人员。

0.2.2 关键术语和Microsoft工具集

商业智能领域充斥着没有正确使用的或者用错的术语。工业上的一些主要的争论主要源自于其他人对于术语的误解，就像哲学上真实的差异一样。记住这一点，我们尝试着保持清晰和一致，即使我们不能解决所有的历史争论。我们将在这里强调一些关键的术语。

当我们定义每个术语时，也强调了关联的Microsoft技术，其中大部分是SQL Server 2005的成员。

- 数据仓库是用于商业智能的平台。在Kimball Method中，数据仓库包含了从原始数据提取到用户见到的软件和应用的所有内容。我们不同意其他作者，他们坚持认为数据仓库仅仅是后备房间的集中式的高度规范化的数据存储，这远离了最终用户。 为了减少分歧，在本书中我们一贯使用短语“数据仓库/商业智能系统(DW/BI系统)”来表示整个端到端系统。当我们特别讨论和专门讨论原子级的用户可查询数据存储时，我们称之为数据仓库数据库。
- 业务过程维度模型是建模数据的特定准则，也是规范化建模的一个选择。一个维度模型包含了和规范化模型一样的信息，但以对称的方式包装数据，这样设计的目的是用户可理解性，商业智能查询性能和对于变化的适应性。规范化模型，有时也被称为第三范式模型，用于支持高数量的、单行插入和更新来定义事务系统，但一般会在可理解性、快速和适应变化方面失败。

我们使用术语“业务过程维度模型”具有两层含义，既指支持业务过程的逻辑维度模型，又指数据库中相应的物理表。换句话说，维度模型既是逻辑的又是物理的。

- 关系型数据库是用于存储、管理和查询数据的一般用途的技术。SQL Server 2005数据库引擎是Microsoft的关系数据库引擎。业务过程维度模型可以存储在一个关系数据库中。支持事务处理的规范化的数据模型也可以存储在一个关系数据库中。
- 联机分析处理(OLAP)数据库是用于存储、管理和查询专门用于支持商业智能使用的技术。SQL Server 2005 Analysis Services是Microsoft的OLAP数据库引擎。业

务过程维度模型可以存储在一个 OLAP 数据库中，但不能存储在一个事务数据库中，除非首先将它变换成明确的维形式。

- 一个 ETL 系统是一个过程的集合，可以清理、转换、合并、重复数据删除、存档，一致化以及结构化数据来用于数据仓库。这些术语在本书中都有描述。早期的 ETL 系统使用 SQL 脚本和其他脚本来构建。尽管许多小型 ETL 系统仍然这么做，但是大型的或者更重要的 ETL 系统是用专门的 ETL 工具。更进一步，几乎每个 DW/BI 系统都使用了像 SQL Server 2005 Integration Services 这样的工具，因为收益很大而增加的代价很低或者代价为零。
- 商业智能(BI)应用是一些预定义的应用，它们可以查询、分析和展现信息来支持业务需要。有一系列的 BI 应用，从复杂的一系列预定义的报告到直接影响事务系统和公司每天操作的分析型应用。可以使用 SQL Server Reporting Services 来构建一个制表应用，以及一个大范围的 Microsoft 和第三方技术来构建复杂的分析型应用。
- 数据挖掘模型是一个静态模型，经常用于根据数据过去的行为预测未来的行为。数据挖掘是一个术语，意指服务于不同目的的统计技术或算法的不固定的(经常改变的)集合。主要包括聚类、决策树、神经网络和预测。Analysis Services 数据挖掘是数据挖掘工具的一个例子。
- 即席查询由用户即时创建，维度模型方法被认为是支持即席查询的最好技术，因为简单的数据库结构易于理解。Microsoft Office，特别是 Excel Pivot 表是市场上最流行的即席查询工具，可以使用 Reporting Services Report Builder 来执行即席查询和简单的报表定义。然而，许多系统为它们的超级用户增加了第三方的即席查询工具来作为对 Excel 和 Report Builder 的补充。
- 此外，数据仓库/商业智能(DW/BI)系统包括：源系统摘要，ETL，既是关系型又是 OLAP 的维数据库，BI 应用以及即席查询工具。DW/BI 系统也包括了管理工具和实践，面向用户的文档和培训，一个安全系统以及其他我们将在本书中讨论的成员。

0.2.3 角色和职责

DW/BI 系统在生命周期中需要一定数目的不同的角色和技能，这些技术和技能可能来自业务和技术领域。在本部分中，我们将回顾和创建 DW/BI 系统有关的主要角色，在角色和人之间一般没有一个一对一的关系。我们和不同团队合作过，这些团队小到只有 1 个人，大到有 40 个人(听说有更大的)，大部分 DW/BI 团队在 3~7 个全职成员之间，并且根据需要可以增加其他人。

单个 DW/BI 团队承担开发和操作的责任是很常见的，这不同于大部分技术项目团队，而且和 DW/BI 项目开发周期的高度迭代是关联的。下面的角色与设计和开发活动相关联。

- DW/BI 经理负责项目的总体领导和方向。DW/BI 经理必须能够和高级业务和 IT 管理进行有效的通信。经理必须能够和团队一起工作来阐明 DW/BI 系统的总体架构。
- 项目经理负责系统开发过程中项目任务和活动的日常管理。
- 业务项目领导者是业务领域的成员并且和项目经理一起工作。

- 业务系统分析师或业务分析师负责领导业务需求定义活动，并且经常参与业务过程维度模型的开发。业务系统分析师需要能够在业务和技术的空白处架起桥梁。
- 数据建模人员负责执行包括数据配置和开发详细维度模型的详细的数据分析。
- 系统架构师设计 DW/BI 系统的不同组件。这些组件包括 ETL 系统、安全系统、审核系统和维护系统。
- 开发数据库管理员(DBA)创建关系型数据仓库数据库并且负责总体的物理设计，包括磁盘排列、划分和初始的索引计划。
- OLAP 数据库设计人员创建 OLAP 数据库。
- ETL 系统开发人员创建 Integration Services 包、脚本及其他可以从源数据库移动数据到数据仓库的其他元素。
- DW/BI 管理工具开发人员负责编写对于管理 DW/BI 系统任何必要的定制工具。这些工具的简单例子包括输入元数据、脚本或执行系统备份和恢复的 Integration Services 包的简单 UI(用户界面)，以及维护维体系结构的简单 UI。
- BI 应用开发人员负责构建 BI 应用，包括标准报告和业务需要的高级分析型应用，他们也负责开发 BI 门户的任何定制的组件以及集成数据挖掘模型到业务操作中。

其他大部分角色也在 DW/BI 项目开发周期的后期阶段起一定的作用，当团队进入部署和操作系统的阶段时，一部分这些角色属于严格操作型的。

- 数据干事负责保证数据仓库中的数据是精确的。
- 安全经理规定业务用户需要的新的用户访问角色。以及增加用户到现有的角色中，安全经理也决定了 DW/BI 系统的 ETL 后备室中的安全过程。
- BI 门户目录经理管理 BI 门户。他决定了门户中的目录以及这些目录如何安排，如何保持最新。
- DW/BI 培训者创建和发布 BI/DW 系统的培训材料。
- 关系数据库管理员(DBA)负责管理关系数据仓库数据库的性能和操作。
- OLAP DBA 负责管理 OLAP 数据仓库数据库的性能和操作。
- 协调经理负责保证 DW/BI 的政策和操作遵循企业和常规的法令，如隐私权、HIPAA 和 Sarbanes-Oxley。协调经理和安全经理和 Internet 审核紧密联系。
- 元数据经理决定了收集哪些元数据、放置在哪里以及如何将它们发布到业务领域。正如我们在第 13 章讨论的，元数据一般不用于管理，除非有专门的人负责。
- 数据挖掘分析师对于业务很熟悉而且常常在统计学有一定的背景、数据挖掘分析师开发数据挖掘模型并和 BI 应用开发人员一起工作，设计使用数据挖掘模型的操作型应用。
- DW/BI 团队的用户支持人员必须能够帮助业务用户，特别是即席查询访问。企业范围的帮助似乎并没有特别必要的技术，只需要帮助解决连接问题。

0.3 本书内容

我们已经将本书分为五个部分：