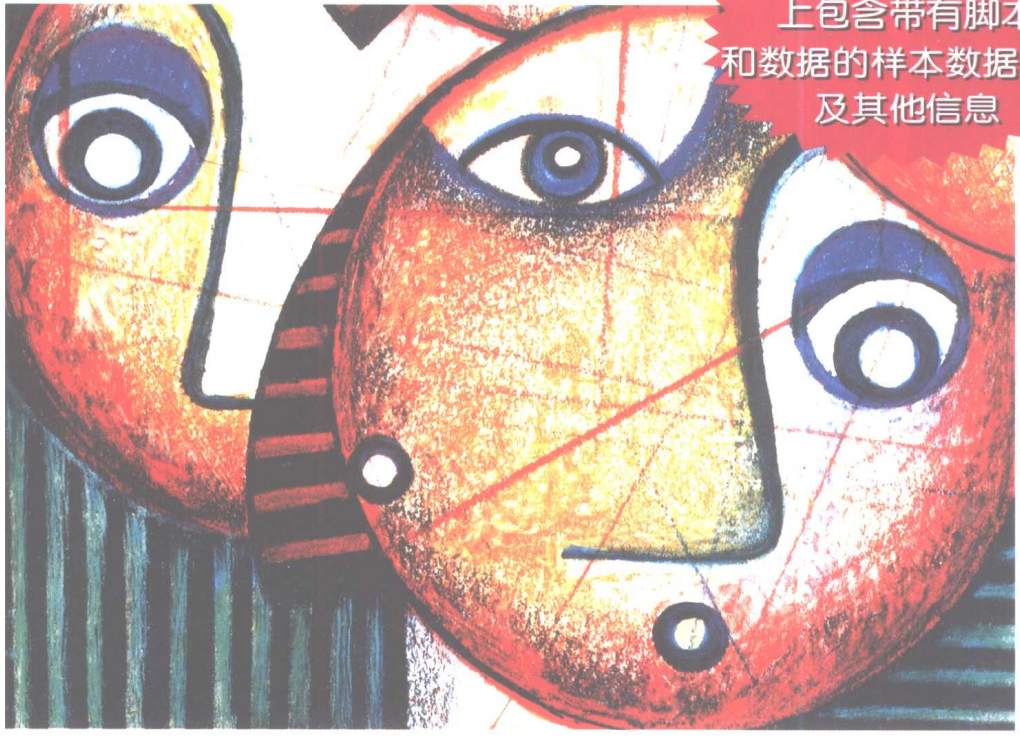
ET PUBLISHING 今日電子

美国IDG“高级开发工具”丛书

Informix/Red Brick数据仓库开发指南

The Official Guide to Informix/Red Brick Data Warehousing

CD-ROM
上包含带有脚本和数据的样本数据中心及其他信息



[美] Robert J. Hocutt 著
伟 峰 等译

借助本书，你将会

- 获取针对数据仓库设计的数据库的最大用途
- 对Red Brick数据仓库的实现有全面的了解

IDG BOOKS WORLDWIDE
PHEI
電子工業出版社
Publishing House of Electronics Industry
URL:http://www.phei.com.cn

美国IDG“高级开发工具”丛书

Informix/Red Brick 数据仓库开发指南

The Official Guide to Informix/Red Brick Data Warehousing

[美] Robert J. Hocutt 著
伟 峰 等译

電子工業出版社
Publishing House of Electronics Industry
北京 · BEIJING

内 容 提 要

本书是一位数据仓库专家的杰作，是多年实践经验的总结，通过浅显易懂的文字深入浅出地介绍了数据仓库的基本知识。包括如何建立数据仓库项目小组和收集业务要求、设计维度模型；将数据组织成更易读取的形式；用Red Brick索引技术和工具减少索引空间和优化索引性能；用TMU与PTMU迅速将数据装入仓库中；用RISQL函数回答SQL无法回答的常见业务问题；充分利用并行处理；用Vista聚集方法提高查询性能。每一章结尾都用“实务经验”总结实际操作中的要点，并通过样本数据库演示实际运用方法。本书是Red Brick用户的宝贵资料，也是学习一般数据仓库知识的优秀教材。

The Official Guide to Informix / Red Brick Data Warehousing by Robert J. Hocutt

图书在版编目(CIP)数据

Informix/Red Brick数据仓库开发指南/（美）霍克特（Hocutt, R.J.）著；伟峰等译.-北京：电子工业出版社，2001.6
（美国IDG“高级开发工具”丛书）
书名原文：The Official Guide to Informix/Red Brick Data Warehousing
ISBN 7-5053-6788-9

Ⅰ.I… Ⅱ.①霍… ②伟… Ⅲ.关系数据库-数据库管理系统，Informix-软件开发 Ⅳ.TP311.138

中国版本图书馆CIP数据核字（2001）第040542号

丛 书 名：美国IDG“高级开发工具”丛书
书　　名：Informix/Red Brick数据仓库开发指南
原 书 名：The Official Guide to Informix/Red Brick Data Warehousing
著　　者：［美］Robert J. Hocutt
译　　者：伟　峰　等
责任编辑：史　平
排版制作：今日电子公司制作部
印 刷 者：北京东光印刷厂
出版发行：电子工业出版社　URL: http: //www.phei.com.cn
　　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：787×1092　1/16　印张: 16.5　字数: 412千字
版　　次：2001年6月第1版　2001年6月第1次印刷
书　　号：7-5053-6788-9 / TP·3817
定　　价：32.00元（含光盘一张）
著作权合同登记号：图字：01-2000-2791

出 版 说 明

为什么要出版这套书

当今世界正处于高速发展的信息时代，网络经济和电子商务彻底改变了人们的思维观念和历史的进程。随着新千年的到来，计算机技术更是方兴未艾。衡量一个国家或一个部门信息化水平的高低依据的是其创新意识和开发能力，特别是计算机软件的开发与应用。目前，国内外各大企业和机构都在不遗余力地竞相开发自己的操作系统和应用程序，使其尽快占领市场，并获取高额回报。为了提高我国计算机软件人员的整体水平，自主开发出国产软件，增强我国的综合实力，使我们在知识经济的大潮面前立于不败之地，我社组织引进了美国IDG Books Worldwide,Inc.出版的这套高级开发系列丛书。

在本丛书的组织翻译中，我们聘请了国内多年从事计算机开发与应用、测试与培训的专家学者，其渊博的知识、丰富的经验，充分体现在本丛书的各个章节中。在翻译过程中，既忠实原著，又充分体现中国文化的特点，而且在技术名词术语、技术内容本身上力求通用、严谨、准确。

这套书的读者对象

本丛书涉及到信息技术方方面面的知识和技能，包括操作系统、硬件配置、开发工具、数据库技术、网络管理、Web设计及电子商务应用等众多领域。因此，本丛书的读者，首先应是一个计算机应用的熟练者和开发者。通过这套丛书，将使您从一名普通计算机用户过渡为专家级用户。也就是说，本丛书的读者对象不是一般的初中级用户，而是具有一定经验，并从事计算机软件开发与设计、系统管理、网络维护等工作的中高级用户。如果您目前就是，或即将成为一名高级开发人员、系统管理员或Web网页的设计者，或许本丛书将会为您提供必要的知识和高级技能。

译 者 序

本书是一位数据仓库专家的专著，是他多年实践经验的总结。很高兴把这部好书翻译过来，献给读者。书中内容丰富翔实，示例生动有趣，都是从实际数据仓库工作中的案例抽象得到的。要想进一步领略、进一步享受，请慢慢翻阅吧。

翻译本书时，正是寒冬初到，但我和张荣、李青、钟锃光、王凌飞一起，在白鹭飞起的地方，怀着巨大的热情挥毫泼墨，细推慢敲，终于拿出自己略感满意的手稿，厦门市电脑学会理事长李棠秋教授亲自审阅了部分章节，刘文琼、刘云昌、刘昌和、严明英、赖华龙、陈凌峰、陈纯颖、周阳生、邹能东、李耀平、彭振庆等朋友也在翻译整理和录排方面提供了诸多帮助。套用一句老话，这是集体智慧的结晶，在此对各位深表感谢。也盼望广大读者不吝赐教，让我们精益求精，更上一层楼。最后，感谢何大曾先生在北京期间为我提供的用车和其他方便。

新世纪已经到来，祝各位读者在新的世纪里大展宏图。

序　言

技术开发常常领先于市场推广，公司需要花大量经费说服用户使用最新产品。而数据仓库市场却并非如此，客户对新的快速装入和快速查询的需求远远超过数据库技术的支持能力。

传统关系数据库针对事务处理，要求快速插入/更新记录和撤销/实现数据完整性。但这样就使传统关系数据库读取数据的速度很慢，花大量资源进行查询，经常与事务处理所需的资源发生冲突，难以管理，很难同时满足OLTP和数据仓库的要求。需要迅速分析数据和所分析数据量的迅速增加引出了全新的行业：开发针对决策支持与业务智能的关系数据库。

我在20世纪70年代后期取得麻省理工大学的管理学硕士学位，然后在一家消费品生产公司工作。我发现学校中所学的大部分量化分析方法都在实际中派不上用场，因为公司并没有保存分析所需的这些信息。虽然能提供某个时段的信息，但很少保存历史信息（即没有趋势信息），现有信息也只能从定期产生的书面报表中取得。对一项800万美元的促销活动进行事前分析的工作中，我在促销活动完成4个月之后才得到了足够信息，发现我们增加了容量，但包装尺寸不合适。我们花了800万美元进行促销，丧失了400万美元的利润，总共损失至少1200万美元。与此同时，市场部也正在计划和执行类似的促销活动。那时候，我的时间95%用在寻找和检索信息，只有5%用在实际分析上。但这种分析终于使管理层体会到保存历史信息的重要性，因此要求事务系统保留档案信息。

第二个主要改变是20世纪80年代初包装产品行业出现的扫描程序数据。在此之前，AC Nielsen和IRI之类的公司在各种产品中提供竞争性销售信息。一种目录可能报告2个月时间内40个市场的800种品牌。利用扫描程序，产品可以具体到40个市场的15 000个项目，在连续层次上提供给这些市场，进行每周和每月聚集。每年管理的不是20万条记录，而是可以增加到100倍，管理2亿条记录。这种数据激增对现有数据库系统的装入、存储和读取功能提出了挑战，业务分析人员开始寻找新的工作方法。

250多家公司选用Metaphor数据存储、读取和视图化系统（现称为IBM的IDS：智能决策系统包）。这些公司的分析人员将数据库设计成符合其业务的思路，他们建立维度表来描述业务的维度与粒度，如时段、产品、市场，并建立事实表，包含量化信息，如销售单位、销售金额和价格。这些维度模型使用户能比Codd和Date建立的传统关系模型（即第三范式和第五范式）更方便地访问数据，因为数据的存放方式设计适合其业务。这种业务的好处是用户可以更方便地查询数据库本身和得到正确的答案。在Metaphor用户群中，我们看到了不同公司在业务维度组中的许多相似之处，这些维度化数据库设计近几年称为星形结构。

设计和实现Metaphor系统之后，我开始为研发Metaphor的Ralph Kimball工作，任顾问职务，和客户一起设计决策支持应用程序和针对查询读取调整数据库。我们发现许多不同行业的数据库设计适合或已经采用维度或星形设计。问题在于，人们还在使用查询性能很差的传统数据库技术。为什么在只需要1500条记录的信息时让数据库读取15 000个页面，且每个页面包含多个记录呢？处理简单查询的时间似乎并不多，但提交的查询一多，系统就受不了。我们可能用到

午餐查询、会议查询、夜间查询、周末查询和各种莫名其妙的查询。我们知道数据库中放有大量数据，有些非常重要，但我们却无法访问。

Kimball发现现有关系引擎通过执行成对连接和用临时结果生成临时表一步一步解析查询。Kimball开始考虑这种优化器技术的好处：先解析维度，然后只对事实表处理一次，消除查询解析中的临时表。由于维度表要比事实表小得多，因此先解决维度表再解析事实表能大大减少查询检索时读取的记录。这个概念是Red Brick数据库引擎的基础。

我们的顾问和工程师小组设计与建立数据库引擎时，对过去使用其他数据库的经历记忆很深。许多人曾经在很少或没有IT支持的条件下管理过自己的数据库，因此我们要让Red Brick易于安装和管理。我们还考虑了可能涉及的数据库负荷，要求数据在查询之前先进行检索。我曾经每月装入26盘数据带，花36小时装入19盘带之后，出现故障，结果又要一切重来。由于快速访问信息和操作可以提高竞争力，因此公司出现这种延迟显然是高成本的、不允许的，甚至危及你的饭碗。

Red Brick的装入器设计成使装入、生成索引和检查参照完整性的过程相当有效。考虑到人们不喜欢用SQL编写应用程序，Red Brick生成了语言扩展，可以方便地进行年度比较（现在在SQL中通过相关子查询支持）、响应排名（前10名）等等。这些Red Brick扩展比新的SQL OLAP扩展提前五六年，在其他数据库预见到这种特殊需求之前，就开创了业务分析的新途径。

Red Brick公司开始开发客户群之后，客户对数据库的赞誉很高，对新特性的要求也很高。我们的第一个保健客户帮我们扩展了维度和结构细节的概念。大型零售商和Internet客户帮助扩展了大型维度表（特别是客户表）如何形成的思想。我们有一个客户是向其他公司提供信息的，他们提出晚间要装入很多数据，而装入窗口很小。如果数据装入发生任何严重问题，则无法实现对用户的数据可访问性的承诺。根据他们的经验，我们重新检查了装入过程中内存、磁盘、缓冲区和多任务的作用。

随着数据仓库与Red Brick Decision Server之类数据库引擎的出现，分析人员可以方便快捷地访问这些数据仓库中的信息，从而将主要精力集中在分析数据、寻找问题和找出对业务有利的解决方案，而不再陷入繁琐和费时的数据定位与检索过程。

对于Red Brick，需求推动着发明的进程。它开始时别的关系数据库厂家还没有数据仓库的概念或认为它只是一种具体应用，它根据用户需求和业务工作的改变而不断演变。你手头的著作是客户长期使用Red Brick的经验总结。《Informix/Red Brick数据仓库开发指南》能帮助你充分利用针对数据仓库设计的数据库。

Mona Pinette
Red Brick Systems公司4号员工

前　言

欢迎使用Red Brick！Red Brick使我的职业开始了全新的历程，它的概念和思想最终成为一流的技术。不仅如此，它还成了做生意和与客户交互的方法。简而言之，就是Red Brick方法。

Red Brick的一部分方法是培训业内人士，从而得以实现我们的技术。尽管如此，技术只是方案的一部分，Red Brick还帮助用户完成工作，也许这是Red Brick最宝贵的财富。因为Red Brick是从顾问公司起家的，我们了解顾问机制，并将这些经验、价值和技巧融入我们的技术中。

这种经验一直沿用到这本Red Brick Decision Server技术的专著。本书提供的知识不仅能在技术上，而且也在一些实现数据仓库所需的“软性技巧”方面帮你以可行的方式取得成功。许多材料是Red Brick（现在的Informix）同事的实际经验和我自己实现Red Brick数据仓库的经验总结。我还提供了很多顾问技巧。

多年来，我介入了大量Red Brick数据仓库项目。从业务分析与建造逻辑模型到实施和性能调整，我在各种情况下和不同应用中采用Red Brick技术。在这些工作中，我发现一条规律：Red Brick中的许多实施问题通常有多个“正确”答案。

了解这个规律将对你的成功大有帮助。如果你想寻找“完美”的答案，则项目永远完成不了。有时，经过认真分析之后，可能只能采用一种可接受的方案，然后继续，这是很正常的。

能否一切“正确”呢？也许不能，但通过利用手头的信息，至少可以保证方向性正确。也许某些组件能做得更好，但这需要经验，需要时才会用到。关键是一次一步，谁都要从头开始，因此，迈开大步前进吧！

读者对象

本书的主要读者是已经或准备实现Red Brick Decision Server仓库者。无论你是项目经理、项目资助者、数据管理员或其他小组成员，只要参与实现Red Brick数据仓库，就应阅读本书。如果你已经安装Red Brick数据仓库，则可以发现书中关于分段、检查和Vista的信息非常有用。

本书也适合顾问使用。每位顾问要支持客户，他们可能在你的组织之内或之外。顾问的客户通常比“客户”本身的客户多，但每个人都可以从书中吸取营养。你会发现，建立Red Brick仓库的技术和收集必要信息的过程有非常密切的关系。

这里的材料按教程形式指导你实现Red Brick Decision Server数据仓库，并介绍许多Red Brick实用资源。因此，许多人都会用到模型和分段的知识。

对读者的要求

技术书籍都对读者的知识背景有一定的要求。读者的类型有多种，有些人要在许多不同情

况下实现大量仓库，例如小组中的顾问。

有些人要实现五六个仓库，而有些人最多只实现一两个仓库。我尽量平衡提供的信息量，照顾不同的读者。虽然众口难调，但我将尽力而为。

本书对读者的要求如下：

- 能访问Red Brick安装——能在Red Brick环境中生成/删除数据库对象和进行各种操作。要学习Red Brick Decision Server如何工作，最好的办法是使用它。建议开发区与生产系统分开，在能建立用户想要的东西之前，先不要破坏用户需要的东西。
- SQL数据库使用知识——Red Brick Decision Server的一大部分是基于SQL。本书不准备介绍SQL及其概念，但我将介绍Red Brick特定的SQL扩展。你应当有一般SQL数据库使用知识，关于ANSI标准SQL的书籍很多，你应先读一本。
- 不能取代Red Brick文档——Red Brick文档是最好的资源，是我见过的最佳文档。本书不想花太多时间重复文档中的内容。如果只是重复文档中的内容，则本书价值不大。至少要阅读所选平台（NT或UNIX）的《Warehouse Administrator's Guide》和《SQL Reference Guide》。如果有兴趣，还可以参阅《RISQL Self Study Guide》。
- 阅读Warehouse Administrator联机帮助——我还假设你熟悉Red Brick GUI管理工具。但是，GUI管理工具没有书面文档。本书中用大量屏幕图形显示管理员如何完成特定功能。维护与管理Red Brick环境的所有工作都可以借助命令行和GUI管理工具进行。

本书内容

本书的目标是介绍如何标识、设计和建立Red Brick Decision Server数据仓库。除了介绍基于Red Brick Decision Server的技术之外，我们还将介绍从客户那里收集必要要求的“软性技巧”。你应按顺序阅读，因为前后文是相关联的。下面列出一些本书内容：

- Red Brick技术组件如何集成
- Red Brick项目基础及如何入手
- 如何确定第一个数据仓库工作
- 如何收集用户要求
- 如何正确开发逻辑和物理数据模型
- 如何在Red Brick中实现模型
- 关于Red Brick索引技术
- 如何使用Red Brick Vista技术
- 如何维护数据仓库

本书将介绍一个样本数据库，介绍在现有Red Brick环境中生成、装入、去掉样本数据库所需的一切。包括如何装入更多数据进行扩充，其中包括DDL脚本、TMU脚本、基础数据和一组产生更多数据的Perl脚本。

本书组织形式

本书分四大部分，大致对应于典型数据仓库项目的流程，帮助确定主要工作段和提供介绍材料的情况。

第一部分　入门

包含第1章、第2章和第3章，介绍选择Red Brick技术、项目基础和项目中要确定的内容，最后详细介绍收集要求。

第二部分　建模

收集要求之后，要绘制表示业务及其工作方法的逻辑模型。第4章详细介绍逻辑建模过程和维度模型的组件，以及如何利用收集的要求建立模型。第5章介绍如何将逻辑模型转变成物理实例，然后将其装入和进行查询。

第三部分　计划实施方案

第6章到第9章介绍如何定义索引、用Vista求值相应聚集表、调整数据库尺寸和实现分段计划。这里有大量信息，此后可能经常引用Red Brick文档。

第四部分　建立与维护数据仓库

第10章到第14章介绍如何安装Red Brick产品（如果还没有安装Red Brick产品）和实际建立数据仓库，还介绍并行查询以及如何装入数据库，最后介绍如何维护Red Brick数据仓库。

此外，大多数章节有两个特殊小节："实务经验"与"样本数据库"。"实务经验"提供Red Brick多年来积累的实务经验，包括销售人员、工程师、技术支持人员和顾问的经验。许多经验教训看起来很简单，但曾让人走过不少弯路。通常，问题就出在一个小地方。

"样本数据库"部分介绍与该章主题有关的样本数据库工作。例如，第9章"分段"中介绍样本数据库的分段计划。

反馈

本书是许多人共同努力的结果。我们努力使内容准确，但错漏之处在所难免。如果发现错误，我们深表歉意，并请读者告诉我们。作为第一次写书的作者，我对读者的意见非常重视，无论是褒是贬。我相信，别人的意见总是有益的。欢迎多提建议和意见。

读者可以登录本书Web站点http://my2cents.idgbooks.com向出版社或作者提意见。

也可以直接和我联系。我不能保证立即答复，但会尽力而为。但希望你的建议具有专业性和建设性。我的E-mail地址为：

rjhocutt@penn.com

目　　录

第一部分　入　　门

第二部分　建　　模

第三部分 计划实施方案

第四部分 建立与维护数据仓库

第一部分

入　　门

第1章 为什么用Red Brick Decision Server

主要内容

- 数据中心与数据仓库
- Red Brick Decision Server技术概述
- Red Brick Decision Server入门
- 实务经验

本章介绍Red Brick Decision Server技术组件和一些数据仓库概念。人们已经介绍了许多关于实现数据中心与数据仓库方面的内容，但目前还没有太多关于用Red Brick Decision Server实现数据仓库的资料。

首先要确定先建立什么——数据中心还是数据仓库。这好像是个政策和市场问题，但实际上是个需要回答的基本问题。这个问题常常引起争议，本章介绍有助于真正可行地解决问题的思路。

然后我们概要介绍Red Brick Decision Server组件，提供其余内容的基线。然后介绍如何寻找潜在项目，包括简要介绍业务运作、资金和其他成功所必需的要素。

本章最后介绍“实务经验”，再次总结本章涉及的主要关键问题。和其他章节中的一样，这些实务经验是从我和我的同事在几百个数据仓库实现方法和多年顾问经验中总结出来的。我希望它有助于你开始自己的数据仓库项目。祝你好运！

1.1 数据中心与数据仓库

未来属于能首先找出并利用业务机会的人，而数据仓库技术对寻找和利用商机至关重要。这种决策支持系统（DSS：Decision Support System）技术提供了理解与管理未来的关键信息。结果，公司可以适应迅速变化的市场条件。在今天的业务环境中（产品周期短、竞争激烈），这样就造就了成与败或小成与大成的差别。

数据仓库的定义多种多样，本书用于决策支持数据仓库/数据中心的实现。决策支持（或联机分析处理，OLAP）的概念提供了业务信息的多维分析。这种多维概念就是按业务规则而不是按数据规则组织数据。详细介绍之前，先要定义数据仓库与数据中心的概念，它们常常引起混乱。

- 数据中心是主题特定数据库，通常由拥有和负责数据的人建立。数据中心是维度的，不经常装入，但有固定计划，可查询性强。
- 数据仓库是整个公司的数据库，跨业务单元。数据仓库可以是数据中心的集合或完全不同的对象，放置整个公司的所有数据，提供面向主题数据中心的数据源。就这么简单。

注意我没有提到数据中心与数据仓库在规模和功能上的差别。这是因为，实际上两者没有什么差别。在最后分析中，只要能解决业务问题，怎么称呼并不重要。根据我的经验，规模与

称呼无关。我见过的最重要和最复杂的数据仓库也不过10GB，我也见过超过4TB数据的数据中心。根据规模判断数据中心与数据仓库是不合适的。

数据中心与数据仓库的选择是学术性的：可以从主题特定的单来源数据中心开始。这个方法有几个好处。第一，如果你首次进入数据仓库世界，则可以先让项目保持简单，学习必要的经验。第二，你花在项目上的时间和经费可能有限，因此应从立竿见影的项目开始。

还要谈谈烟囱管（storepipe）数据中心。这种数据中心是建立在某种“真空”中的，每个数据中心相互不知道，项目小组相互封闭，关键决策以不一致的方式进行。

这种思想使许多人未能从正确的地方入手，最终只好放弃或发现错误和吸取教训。如果建立烟囱管数据中心，则项目小组和项目资助者有责任了解别人的工作，并采取相应措施。

都是为了用户

数据中心与数据仓库的术语混乱（在我看来）主要是不同公司争夺市场份额的促销活动造成的。这些特征反映的不是实际做法，而是技术方面的差别。这样虽然不错，但经过一段时间后，市场消息通常会掩盖问题的实质。

人们通常认为搞不清楚的或自己不懂的东西都是不好的，整个数据仓库/数据中心的问题也是这样。许多软件的销售仅仅基于人们的担心，建立数据仓库或数据中心不是为了解决技术问题，而是为了用户。

不要忘记问题的实质，数据仓库从用户开始，最终要着眼于用户，让他们及时回答常见的业务问题。在最后分析中，用户并不关心你用什么技术、如何设计以及其他任何技术问题，用户只要知道何时能够得到、了解和方便地使用。

数据仓库是现有技术在处理一些业务报表问题时捉襟见肘的条件下产生的。你有两种选择：可以坐而论道，把数据仓库与数据中心问题争个面红耳赤；也可以做一些实事，为解决用户业务问题实行一个方案。实际上，怎么称呼并不重要，只要实用就好。

1.2 Red Brick Decision Server技术组件概述

Informix Red Brick Decision Server是为数据仓库、数据中心和联机应用程序处理（OLAP：On-Line Application Processing）应用程序设计的关系数据库管理系统（RDBMS：relational database management system）。Red Brick Decision Server中的许多技术组件多年来一直是独此一家，尽管新的数据库已经迎头赶上，但至今还有一些组件是只在Red Brick Decision Server中存在的。

许多负责技术组件的人都曾经是其他数据库中的行业领头人，如Teradata、Oracle和DB2 Parallel Edition。他们的一流技术造就了Red Brick Decision Server的今天。下面简要介绍Red Brick Decision Server中的主要技术组件，让你对其集成方法有基本的了解（后面章节将提供详细信息）。

1.2.1 STARjoins与索引

维度模型包括事实表（保存业务的量化数据，即要查询的事实）和维度表（保存描述维度属性的数据）。维度表通过外部关键字引用连接事实表，如图1.1所示。

图1.1演示了维度表通过外部关键字引用连接事实表，这些表来自样本数据库。Billing_