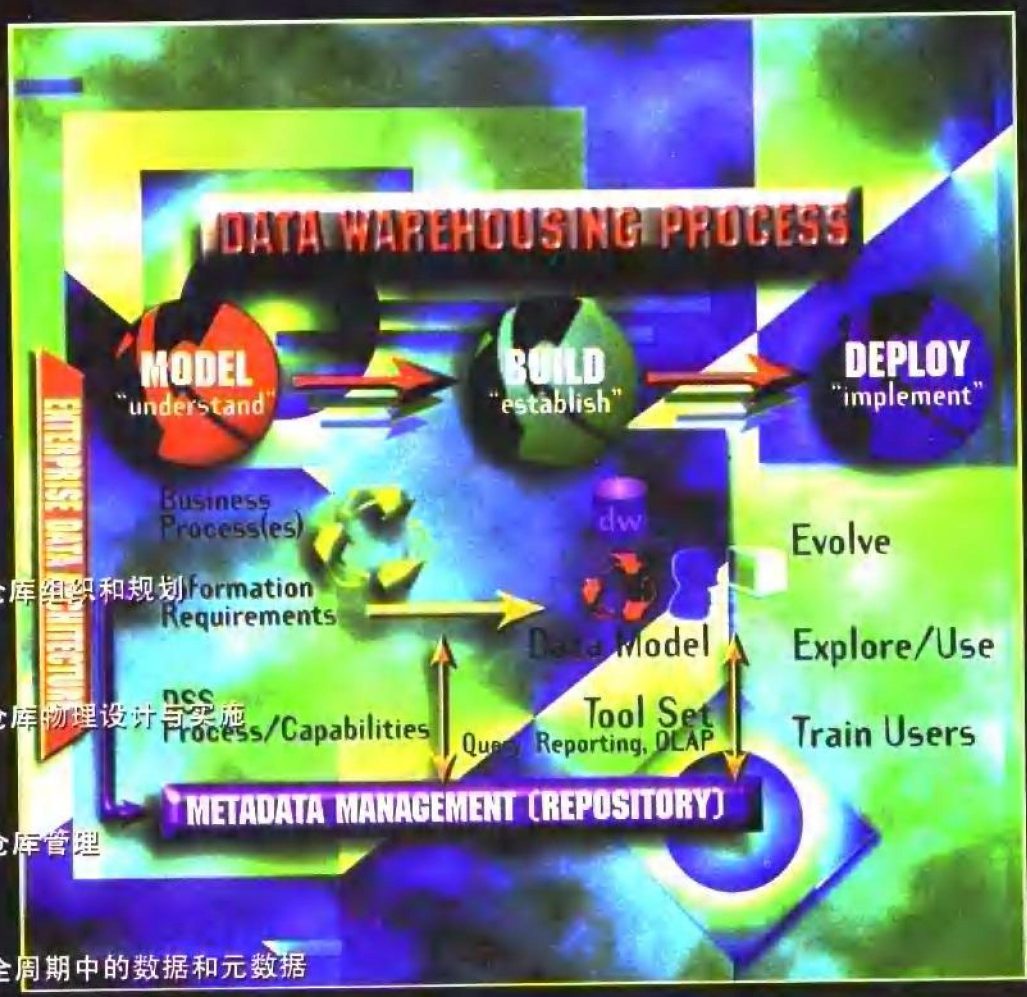
DATA WAREHOUSE
PRACTICAL ADVICE FROM THE EXPERTS
数据仓库技术

IBM / Lotus 软件技术系列丛书

〔美〕Joyce Bischoff　Ted Alexander 著　　成栋、魏立原 译　　周生炳 校



- 数据仓库组织和规划
- 数据仓库物理设计与实施
- 数据仓库管理
- 管理全周期中的数据和元数据

電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
URL: http://www.phei.com.cn

IBM/Lotus 软件技术系列丛书

数据仓库技术

[美]Joyce Bischoff　Ted Alexander　著
成　栋　魏立原　译
周生炳　校

電子工業出版社
Publishing House of Electronics Industry

约翰 · A · 查赫曼: 加州格伦代尔市 Zachman International 总裁

关于作者的详细信息及联系方法，可参见本书后的“作者简介”。 我们还要感谢 Prentice Hall 出版社的编辑阿伦 · 亚普特、沙林 · 约特拉诺、玛丽 · 安 · 泰拉特尼克和劳拉 · 斯蒂尔对本书始终不渝的支持。我们同样要感谢所有的作者、审稿人和其它给予我们建议和支持的人们。除这些作者以外，要感谢下列人士:

IBM 公司的查克 · 巴拉德

Bischoff Technical Services 公司的查克 · 毕晓夫

Information Bridges 公司的伊丽莎白 · 勃特基维奇

E. I. DuPont de Nemours 公司的戴维 · 克里斯蒂安

IBM 公司的杰瑞琳 · 葛兰维尔

Responsive Systems 公司的乔尔 · 戈德斯坦

Strategic Frameworks 公司、加州大学欧文分校的伯尼 · 杰尔特玛

IBM 公司的戴维 · 里德尔

Ford Electronics 公司的罗伯特 · 马尼里

IBM 公司的罗杰 · 米勒

General Accident 保险公司的安妮 · 玛丽 · 史密斯

我们还要感谢不计其数的专业人士，他们通过写作和演讲与我们分享着数据仓库的经验。最后要感谢我们拥有过与许多客户分享数据仓库的机会，他们都曾直接或间接地对本书作出贡献。

小结

本书是为在其组织开发数据仓库时起关键作用的读者而写的。虽然每个行业中都有数据仓库，但有些数据仓库的开发明显成功；有些起初看上去很成功，但后来由于种种原因还是失败了。有些则一开始就是失败。随着数据仓库的发展，必须考虑许多组织结构的变化、技术的变化和管理的变化。各位作者将在下文中与大家分享其丰富的经验。我们希望读者能从中找到通向成功数据仓库之路的方针策略、技术指南和方向。

乔伊斯 · 毕晓夫

泰德 · 亚历山大

内 容 简 介

正如本书副标题所言，这是一本由数十位长期从事数据仓库技术的国际专家的实战经验结晶而成的巨作。其目的在于帮助技术经理、项目经理、业务分析员、系统设计员、数据库管理员、系统管理员、最终用户等规划、设计、开发、实施和管理数据仓库。本书所包括的六大部分阐述了专家们对实现以上宗旨的方略与见地，讨论了企业设计大型数据库、实施安全措施和管理数据仓库以实现数据仓库的整个过程中各种各样的问题。论述了在复杂多变信息时代中企业的实际经营和战略。引进本书旨在帮助和推动我国数据仓库技术的发展，发挥数据仓库技术在现代化企业中的推动作用，构筑我国高效、实用、科学的数据仓库。本书最适用于计算机科学、信息系统和工商管理专业高年级学生使用，本书值得每位信息专业和管理界的人士阅读，无论他/她是否从事着数据仓库工作。

丛 书 名：IBM/Lotus 软件技术系列丛书

书 名：**数据仓库技术**

著 者：[美]Joyce Bischoff, Ted Alexander

译 者：成栋 魏立原

校 者：周生炳

责任编辑：操龙兵

特约编辑：张德强

排版制作：电子工业出版社计算机排版室

印 刷 者：北京京安达明印刷厂

出版发行：电子工业出版社出版、发行 URL：http://www.phei.com.cn

北京市海淀区万寿路 173 信箱 邮编 100036 发行部电话：68214070

经 销：各地新华书店经销

开 本：787×1092 1/16 印张：18.75 字数：464 千字

版 次：1998 年 6 月第 1 版 1998 年 6 月第 1 次印刷

书 号：ISBN 7-5053-4562-1 / TP·2151

定 价：35.00 元

版权贸易合同登记号：图字：01－98－0110

序

本书的副标题是“数十位国际专家的经验结晶”，而它的确名副其实。我还从未读过一本汇集了这么多经验的书籍。本书由多位作者以各种风格从不同的专业领域和角度论述了多种专题，但贯穿始终的是每章都透着的权威的口吻，这是那种毫发无损地处理了可能的灾难后的沉着和取得巨大成功后的喜悦中所体现出的权威。

本书各章的作者都实际参与了数据仓库的实现并取得了成功，了解对成功至关重要的因素。本书值得每位信息专业人士阅读，无论他是否从事着数据仓库工作。这是一本论述在复杂多变的信息时代中企业实际经营状况的书籍。

本书的坦率和诚实给我留下了深刻印象。编辑和作者们并不投你所好，告诉你顺耳之言；他们为你提供的是经验，使你能享受到他们来之不易的成功所带来的益处，告诉你采用数据仓库技术所需要的知识。如果你采纳了这些建议，那么无论何时何地实现数据仓库，你一定会发现这些概念的宝贵价值。

虽然我（还有其他许多人）是从我的好朋友比尔·因蒙那里知道数据仓库的概念，但我想同时还有其他人也“发明”了这个概念，因为数据仓库这个观念的时代已经来临了。托马斯·库恩的《科学革命论》是本有代表性的著作。在这本书中，库恩认为发明创造是时代的产物，而不是发明者个性的作用。同样，彼得·陈、查理·巴赫曼、鲍勃·布朗和克里夫·芬克斯坦（我想还有一些我不知道姓名的人）都同时发明了数据模型的概念。而数据仓库的确是这个时代的产物。

数据处理的历史里一直盛行的观点是：企业应用中计算机的价值在于计算机较之人类而言能够更快更精确地处理重复事务。因此，大多数系统都是面向联机事务处理（ OLTP ）。而数据的价值就在于保证完成一项事务，所以数据的体系结构一直被看作非常无聊的话题，既耗时又费钱。

只有当信息时代企业环境的复杂性和动态性开始让人们强调管理者对企业内外所发生变化的洞察力时，数据的价值和语义质量才开始变成一个业务问题，而不再是一个技术问题。这时，以前系统应用中所遗留的不一致性和不连续性（即缺少数据结构）就成为信息社会的瓶颈问题，使企业难以抑制对其应用系统的挫折感，同时也增加了系统维护的时间和费用。

这种情况下，就提出一种弥补数据体系结构缺陷的办法，即建立一个虚拟的集成数据库，存储现有的真实历史数据，从而尽可能降低物理的不一致与语义的不连续问题，使现有的数据可应用到单纯事务处理之外的管理目的。数据仓库就注定要出现了。

我之所以描述历史上数据仓库出现的来龙去脉，是因为那些正在建立数据仓库却忽视数据体系结构问题的人显然没有真正理解这个问题。如果数据仓库不是源于企业的数据模型，那么它只是又一份历史文档，最终会进一步增加企业对其信息系统的失望和挫折感。

在阅读本书时，我发现每位作者都建议读者先建立一个逻辑数据模型，而且要先了解这个模型的业务含义——即数据体系结构，然后再从历史文件中提取数据来建立数据仓库。这就充分说明本书充满了真知灼见，能使你成功地将这些重要的概念成功地应用到企业中；而不是不顾它们的业务和技术的有效性和实用性，得过且过地予以实现。

数据仓库有两个优点。一方面，它使我们不必重新编制输入（事务）处理系统，就能够建立一个结构化的环境，将输出（决策支持）处理移入新环境（数据仓库）里。另一方面，它建立概念模式（逻辑数据模型）、内部模式（历史文件）和外部模式（数据仓库）的三模

式环境，其中，历史文件的“多对多”（m × n）维护关系可简化为三模式环境下的“多对一对多”（m + n）的关系。重要的是，要注意只有当逻辑数据模型是数据仓库的必要组成部分时，这两个优点才会显现出来。

除了这些短期利益外，数据仓库发展的下一步将是逐个事务地重建输入（事务）处理系统，以支持和直接满足结构化的（业务数据商店）环境；进而还将逐个系统地消除历史文件所有内在的不连续性、维护问题和成本方面的缺陷。然而，这就意味着，建立逻辑数据模型的目的是为了获得事务处理物理数据库（业务数据商店）和决策支持物理数据库（数据仓库），并且保证两者的实现始终同步。

因此，数据仓库的直接价值在于满足管理者对于全面了解企业内部情况和外部环境的短期需要，同时减少现有系统的维护负担。在向纲目库过渡的过程中，这对于降低压力和改善管理者和信息技术人员的合作方面都是非常有价值的。

另外，数据仓库的长期价值在于通过业务数据商店使企业能够从历史系统转移到适应企业高速变化的结构化环境。由于企业的生存依赖于对市场变化的反应能力，所以这对于信息时代中的企业有着极其重要的意义。有人曾雄辩地论证说，新企业（竞争者）、规则改变、现有的竞争和市场总需求等概念将迫使人们重写历史文件。这只是个时间问题，但问题在于只是重写历史文件，还是使其结构化以适应变化呢？我认为，数据仓库是结构化和进入新环境的关键所在，而体系结构又是数据仓库和适应高速变化的关键。

无论是着眼于数据仓库的直接短期益处，还是同我一样着眼于其长期益处中的重大意义，负责的建议都是基于企业数据模型来建立数据仓库。

《数据仓库：数十位国际专家的经验结晶》一书中的所有章节的作者都是有真知灼见的、有责任心的人。它是对成功地实现了数据仓库的专家的智慧和经验的颂扬和证明；其中讨论了从设计大数据库、实施安全措施和管理数据仓库的企业实现数据仓库整个过程中各种各样的问题。

严肃审慎地来说，近来数据仓库的主题是很脆弱的。很难找到一份行业出版物或开发商的说明手册中没有大肆宣扬数据仓库及其在解决企业所有信息问题的潜力。这不禁让人回想起 CASE、Repository、ADCycle、甚至更早以前的人工智能、专家系统和小型计算机等当年的遭遇。CASE、Repository、ADCycle 和所有其他技术本身没有问题，但任何技术一旦被当成包治百病的灵药，就会立即发现它们无法满足那些过高的不现实的期望。当一颗“银弹”暴露出缺陷之后，那种典型的强烈对抗性的市场反应足以摧毁一切，甚至包括其中最基本和最关键的概念。我感觉数据仓库会成为 1997 年的包治百病的良药。“那东西不管用！”的反应是实实在在的现实，是使我们免于重蹈覆辙的好方法，它将有助于取得更大更卓越的成就。

《数据仓库：数十位国际专家的经验结晶》一书的确反映了现实的情况，我奉劝你们认真地阅读，采纳其意见，并取得巨大的成功。我们每个人都必须确保数据仓库方法的生命力和可靠性。

本书无疑是乔伊斯·毕晓夫和泰得·亚历山大深思熟虑和艰辛努力的结晶。他们两人汇集了对数据仓库技术目前水平的广泛而实际的评述。我们不妨几年后回过头来看一下，那时会发现，他们在保护这些珍贵概念的完整性方面为未来的信息工作者做出了重大贡献。

约翰·A·查赫曼

于加州格伦代尔

前言

信息技术的进步已使我们到达实现最终信息目标的边缘：即以合理的价格及时地向任何地点的任何需要数据的人提供高质量的数据。信息技术的发展已经把我们从简单的批处理和联机事务处理的时代带入了数据仓库、联机分析处理（OLAP）、Internet 和 Intranet 的时代。这段旅程转瞬即逝，其中却充满了先驱者（往往也是流血者）的痛苦和艰辛。

这本关于数据仓库的手册的目的在于帮助技术经理、项目经理、业务分析员、系统设计员、系统体系结构设计员、数据库管理员（DBA）、数据管理员（DA）、系统管理员、最终用户和数据仓库项目小组成员规划、设计、开发、实施和管理数据仓库；这本书也适于计算机科学、信息系统和工商管理专业的学生阅读；数据仓库产品的开发商也会发现本书对培训其员工和客户实施数据仓库的方法方面的价值。它有助于人们作好充分的准备，朝着随时随地提供任何信息的目标不断进取。作者们都是来自不同公司的顾问和用户，他们都实施过数以百计个数据仓库；他们会与你分享他们的经验，使你避开实施过程中的陷阱。

本书的使用

这本书包括六个部分，这样安排是遵照典型的数据仓库项目实施过程中管理人员和项目小组所需要的信息的大致顺序。建议每位读者阅读本书的前三部分，以获得数据仓库的基本认识。尽管读者们可以从头到尾读完全书，但实际上没有特定的阅读顺序，读者可在处理数据仓库项目某方面问题时查阅本书*。

如果将本书用于教学，它最适用于计算机科学、信息系统和工商管理专业的高年级使用。这门课程强调了使用信息是为了满足企业的目标，并谈到了满足这些目标所涉及的业务和技术问题。计算机科学和信息系统专业可从头到尾阅读本书；工商管理专业可阅读第一、二、三、六部分，如果学生具备技术背景知识，也可有选择地阅读第四和第五部分。许多大学都为信息技术领域人员开设进修课程，这本书也是这类课堂的理想教材，可从头到尾阅读。

每章结尾有参考书目，在文中以上标数字标注。

第一部分：数据仓库初步

这一部分是理解建立数据仓库所涉及问题的基础。它建立了数据仓库的观念，讲述了数据仓库要遵循企业的目标。它评价了开发各阶段中组织的影响和与最终用户的合作的必要性。有经验的用户提供的建议对正考虑实施数据仓库项目的人很有价值。第一部分对于技术经理、项目经理、业务分析员、信息资源管理部门和最终用户都大有裨益，即使尚未决定实施数据仓库。

第二部分：数据仓库的规划

别人都是如何建立数据仓库的？这种方法与联机事务处理所用的方法真的不同吗？我们应当如何规划数据仓库项目？如何考虑其成本？这一部分讨论的是数据仓库的开发方法，探讨了与试验项目规划有关的问题。项目经理、设计人员、开发人员和管理者可从中了解方法论的差异和数据仓库开发的反复性。

---

* 为了方便阅读，本书未提供了缩略词表

第三部分：数据：关键的问题

数据是数据仓库的基础。第三部分先研究数据质量，接着又讨论了数据仓库中元数据的问题。它讨论了一家大银行的面向用户数据目录的开发经验，这对希望开发自己的目录或购买商业化产品的人是很有价值的。由历史环境转到数据仓库时数据的分析和转换是人们关注的焦点，因为这项工作几乎占开发数据仓库一半的时间。这一部分是技术经理、项目经理、业务分析员和信息资源管理部门的兴趣所在。

第四部分：数据仓库的设计和实施

在设计之前必须了解数据仓库的体系结构，接着是制定数据的体系结构和物理设计。这一部分讨论了技术设计方案，包括数据复制、联机分析处理（OLAP）、中间件和并行技术。这一部分会吸引数据库设计者、应用程序的设计者和管理者的注意。

第五部分：数据仓库的管理

许多数据仓库是在异质硬件和软件环境下实现的，这就带来了许多困难。这一部分讨论复杂环境中必须加以控制的管理问题。这正投合了数据库管理员和系统管理员的特殊兴趣。

第六部分：数据仓库的发展趋势

数据仓库技术的发展趋势是什么？对试图为未来奠定基础的经理、业务分析员和项目小组成员而言，这部分是十分引人入胜的。

致谢

我们要感谢所有作者，通过他们对这本书的贡献把专业知识与大家分享。我们还要感谢他们所有其它的评论、建议以及其他形式数不胜数的支持。

希德·阿代尔曼：加州谢尔曼欧克斯市 Sid Adelman and Associates 公司

彼得·布鲁克斯：麻省列克星顿市 Coopers and Lybrand 咨询公司高级技术部　　玛丽·布莱塔：新泽西州普林斯顿市 Marie Buretta 公司总裁　　布兰达·卡斯蒂尔：加州洛杉矶市 Ernst and Young 公司高级经理　　霍华德·福斯迪克：依利诺伊州维勒帕克市 Fosdick 咨询公司

苏珊·高斯登：英国萨里郡韦布里奇县 Brooklands Technology 公司董事

戴维·格里森：PLATINUM Technology 公司属下的 PLATINUM 信息管理咨询公司咨询与产品高级经理　　保罗·海辛格：佐治亚州亚特兰大 Vision UnlimITed 公司管理董事

马丁·胡贝尔：加拿大多伦多市 Martin Hubel Systems 咨询公司

丹尼斯·科萨：纽约州纽约市 Chase Manhattan 银行企业信息体系结构部副主任

约翰·拉德雷：密苏里州圣路易斯市 META Group 公司

特里娜·拉鲁：新泽西州萨默塞特市 AT&T 公司高级顾问

杰伊·马奎斯：佐治亚州亚特兰大 Praxium Group 公司高级合伙人

特晨·梅森：英国萨里郡韦布里奇市 Brooklands Technology 公司董事

爱德华·M·彼得斯：Intersolv 公司 DataDirect 分公司副总裁兼总经理

杰克·斯威尼：麻省温思罗普市 Intellidex Systems 总裁兼首席执行官

匹特·尤罗齐克：IBM 公司圣特里萨实验室高级技术委员会委员

科特·怀特：加州摩根希尔市 DataBase Associates International 公司

理查德·耶维奇：佛罗里达州基比斯凯恩市 RYC 公司经理

目录

第一部分　数据仓库初步

千里之行，始于足下。——老子

1、	数据仓库导言	乔伊斯·毕晓夫
2、	信息技术的复兴	保罗·海辛格
3、	组织和文化	希德·阿代尔曼
4、	与最终用户有效地合作	乔伊斯·毕晓夫
5、	七宗罪	丹尼斯·科萨
6、	实际的数据仓库：管理的挑战	霍华德·福斯迪克

各行各业中都有许多数据仓库取得了不同程度的成功。而不成功原因往往在于对数据仓库的基本问题没弄清楚。这一部分中，有经验的用户和咨询顾问提出了针对建立数据仓库最初阶段的意见。

第一章“数据仓库导言”的作者是乔伊斯·毕晓夫。这一章阐述了整体信息战略、定义了数据仓库的基本概念和术语，为理解数据仓库技术奠定了基础。

第二章“信息技术的复兴”的作者是保罗·海辛格。这一章提出了一种充满幻想和激情的管理框架。海辛格讨论了信息技术行业目前的复兴；指出要利用数据仓库发展的潜力，头脑也需要复兴。他强调了数据仓库战略遵从企业目标的必要性，指出了成功的关键因素，并勾勒出实现管理目标的行动计划。

第三章“组织与文化”的作者是希德·阿代尔曼，讨论了一个成功的数据仓库必须解决的问题，阿代尔曼谈到了建立一个合适的基础结构以支持仓库的问题，并列出了成功数据仓库的十条标准。他描述了不断变化的角色和责任以及在仓库实施中权力结构的变化。

第四章“与最终用户有效地合作”的作者是乔伊斯·毕晓夫，进一步探讨了组织问题，许多仓库失败的原因在于用户没有适当地参与实施各个阶段。项目的每个阶段中最终用户的参与水平会因仓库项目类型的不同而不同。毕晓夫建议让用户参与仓库的规划、设计、实施和管理。有一件事是肯定的，即，如果用户只在仓库实施之后才参与进来，仓库就很可能遭到失败。

在第五章“七宗罪”中，丹尼斯·科萨列出了一系列在建立仓库时不能做的事。这些建议都来自于他在一家大型的纽约银行建立仓库的实际经验。

在第六章“实际的数据仓库：管理的挑战”中，霍华德·福斯迪克分析了数据仓库项目失败的最常见原因，并且列出了数据仓库项目常犯错误的清单。

第一章　数据仓库导言

特拉华州霍克欣市 Bischoff 咨询公司　乔伊斯·毕晓夫

没有人能看得很远，多数人都是鼠目寸光。——托马斯·卡莱尔

必须考虑作为数据仓库的业务基础的基本战略。数据仓库开发过程中有四个层次的分析处理。数据仓库自身并不是终点，只是向信息数据超市又进了一步。

信息的战略供给

由于对随时随地访问任何信息的要求是永无止境的，就需要一个体系结构来容纳各种格式的内部和外部数据，其中必须包括有经营数据、历史数据、现行数据、订阅数据库及来自 Internet 服务商的数据，还必须包括易于访问的元数据。今天的企业要求访问并综合来自各种数据商店的数据，能够对其进行复杂的数据分析，并创建代表业务分析员对数据的看法的多维数据视图。而且还需要概括、细化展示、多层次/多视角查看跨主题和业务范围的信息。

问题

这些目标很难达到，因为数据是分散在多种互不兼容的结构中。文档的缺乏妨碍我们把旧系统和新系统集成起来。虽然数据库管理员一直在倡导文档完善的、集成的数据库设计，使其能为各种应用程序共享；但大多数组织并未重视这种意见。结果使分析员不建立集成的数据结构就无法进行跨应用程序的分析。

Internet 提供了新的业务机会，同时需要非传统的信息类别。 WWW 为所有人提供了杂志、旅行、专业沟通、游戏、购物信息和整套全新的活动。开发商正在把多种来源的数据合并成数据库，销售使用 Internet 对数据库的访问权。虽然数据及其关系并未在数据目录中进行标准化和定义，但已开发出用来识别相关信息的检索引擎。对同一检索目标不同检索引擎会产生不同的结果，这就意味着需要改进的 Internet 软件和跨组织的更精确的元数据。

分析处理的需求

在现代组织中，信息系统需要支持至少四个层次的分析处理工作。第一个层次是对当前数据和历史数据的简单查询和报表，通常由电子报表、查询工具和报表生成工具实现。第二个层次则深入进行跨数据商店的“what if”处理。如果第二年劳动力成本增加 5%，而销售额不变，那么利润会怎样？使用电子报表、查询工具和数据库技术就能回答这个问题。

第三个层次分析过去究竟发生什么事件使得数据呈现目前状态。为什么一月份里东北地区的咳嗽糖浆销售量突然提高，而其他地区却没有变化呢？是因为东北地区进行了特别有效的营销活动吗？是因为那里流感蔓延吗？还是因为哪个竞争者退出了？要回答这些查询要求，就得对内部数据和外部数据进行复杂的处理。这一层次的查询活动中对外部数据的需求日益增长。信用卡公司希望在某家大公司裁减雇员时了解潜在的失业状况，失业情况会影响持卡人的支付能力，从而增加坏帐数额。目前还没有技术能够支持这类自动分析。

第四个层次要分析过去发生什么事件和未来应完成什么工作才能实现某种特定改变。例如，如果明年利润目标是增长 5%，哪些前提条件变化才能达到这一目的？此外，还须采取哪些措施来实现这些变化？目前还没有技术支持这类分析工作，开发商正努力提供支持这种需求的产品。人工智能应用于非电子游戏的产品来提高企业竞争力的时机已经成熟。

信息数据超市（超级数据仓库）

组织应该开始转移到最终支持对各种格式的内、外部数据进行各个层次的访问和分析处理的体系结构上。“信息数据超市（ IDSS ）”是毕晓夫和耶维奇[1]在一篇论文中提出的，指的是满足在四个层次上对内、外部数据进行分析的全面需求所需的体系结构，包括访问历史系统、业务数据商店、数据仓库、数据商场、数据库服务供应商及 Internet 与 Intranet 的网站服务器。 IDSS 也可称作“超级数据仓库”，因为它远远不止是目前的数据仓库。

虚拟仓库技术能提供对存储在任何位置的数据的访问，并通过为最终用户屏蔽物理位置的语义映射提供跨数据商店的视图。促进互可操作性的行业标准也将允许应用程序和数据库间的互可操作性，而且随着 IDSS 的不断发展，还会得到更广泛的应用。

IDSS 体系结构主要着眼于使用所有数据商店后分析处理所发生的重大变化。除了前面提到的分析的一般需求外，用户可能还会提出一些只有对多个业务组织进行异种联合才能回答的问题。例如，就某一特定信用等级来说，一笔 20 年期固定抵押率的抵押业务的最低抵押率是多少？在某一顾客住宅的 100 英里范围内，何处可买到最便宜的奔驰车？

虽然目前的数据或技术的协同还不能够实施完整的 IDSS ，但开发商还是努力提供有效的产品。将有共同的含义但定义略有差异的各个字段上的数据联合起来，便是一种需求。例如，社会保险号在一张表中用了破折号，在另一张表中用了空格，在第三张中则是 9 位数值型字段。需要采用模糊逻辑或人工智能来识别这些其实是一样的字段，还需要修改结构化查询语言（ SQL ）以支持这类处理。为标准化和集成每个文件，尤其是组织之外的文件，而进行的不断转换格式的繁重负担早已让人不堪忍受了。虽然理论上说这可能在组织内部实施，但不可能控制外部数据库。

另一种缺乏的产品是用来支持跨组织数据访问的动态的、主动的、非结构化的数据目录。方法之一是将关键的、面向顾客的元数据移入到关系或后关系（面向目标的）数据库管理系统的系统目录中。关系目录的底层存储了对每张表、表中每列、来源和关系的基本描述。这就满足了 E.F. 科德[2]提出的关系理论，也使用户或应用软件不须使用具体的目录就能直接访问数据。它还能使每个开发商的数据库或目录产品都能访问标准格式的数据，并将其与特定的开发商产品的扩展功能集成起来。这就要求关系数据库管理系统（ RDBMS ）的开发商与数据库/数据字典的开发商进行合作。一些数据库/数据字典的开发商也许认为这是对他们领域的侵犯，但事实并非如此。提供丰富功能的目录开发商的地位不仅不会动摇，而且还会有所增强。他们可以继续使用自己专用的数据结构，只在 RDBMS 中存储关键的面向最终用户的信息。

另一个主要的变化是在关系或后关系系统中存储业务规则和约束的方式。多年来程序员一直在应用程序编制执行业务规则和约束的代码。而开发人员和管理人员现在已开始将业务规则和约束加入数据库的定义中，由数据库引擎来保证这些定义，不必再由应用程序来完成。这种做法改进了数据的完整性和一致性，保证了这些规则能得到始终贯彻。

用于减轻数据管理员（ DA ）和数据库管理员（ DBA ）负担的管理工具的需求也很大。各种仓库工具所支持的维护多数据库中元数据的工作需要丰富的专业知识和大量的管理工作。 Metadata Coalition 的功能是减少各工具间数据的交换，但对数据移动的管理仍然是个难题。可在各产品中加入触发程序来自动更新其他库。如果物理数据商店的管理采用在意外发生时能通知 DBA 的智能代理程序的话，就可进一步减轻 DBA 的负担。

IDSS 体系结构还引入了称为 IDSS 用户访问层（ UAL ）的新层次，它能压缩访问代码，并允许 IT 人员对其进行管理和优化。元数据管理员可通过 IDSS 的 UAL 解决数据格式差异问题和数据位置问题。这个访问层既可作为简单的通路，也可充分参与。必要时会与元数据管理员进行“讨论”来解决位置或数据问题。这就是 IDSS 的真正核心。就象在服务器上集中

了描述代码和过程代码（如存入的过程和触发程序）的客户机/服务器结构一样，可将访问代码从客户机上去除。这就使 IT 人员能集成动态与解释处理的新方法，并充分发挥专家系统的优势。

图 1.1 所示为用户角度的 IDSS 视图。注意，图中所示的各种概念也可在传统的服务器、Intranet 服务器或 Internet 服务器上实施。

虽然今天的技术还无法全部实现 IDSS 的长期目标，但现在应该制定策略来建立 IDSS 的基本结构。这些结构应该提供数据的完整性、灵活性、可伸缩性、可用性和使用的便利性。在变化无常的环境中应使用稳健的关系或后关系数据库管理系统，既坚持了行业标准，又保证了数据的完整性。虽然从小开始来慢慢发展总错不了的，但公司在选择硬件、软件和应用程序时应规划好可伸缩性。今天，数据仓库在前六个月里扩展了一倍是很平常的。当数据大量增加时，如果发现应用程序和支持软件要超出其技术极限的话，那么想包容的用户数和数据量的增加会耗费极高。连接问题取决于行业标准。 IT 部门必须保证有必要工具来连接整个企业的客户机/服务器系统。

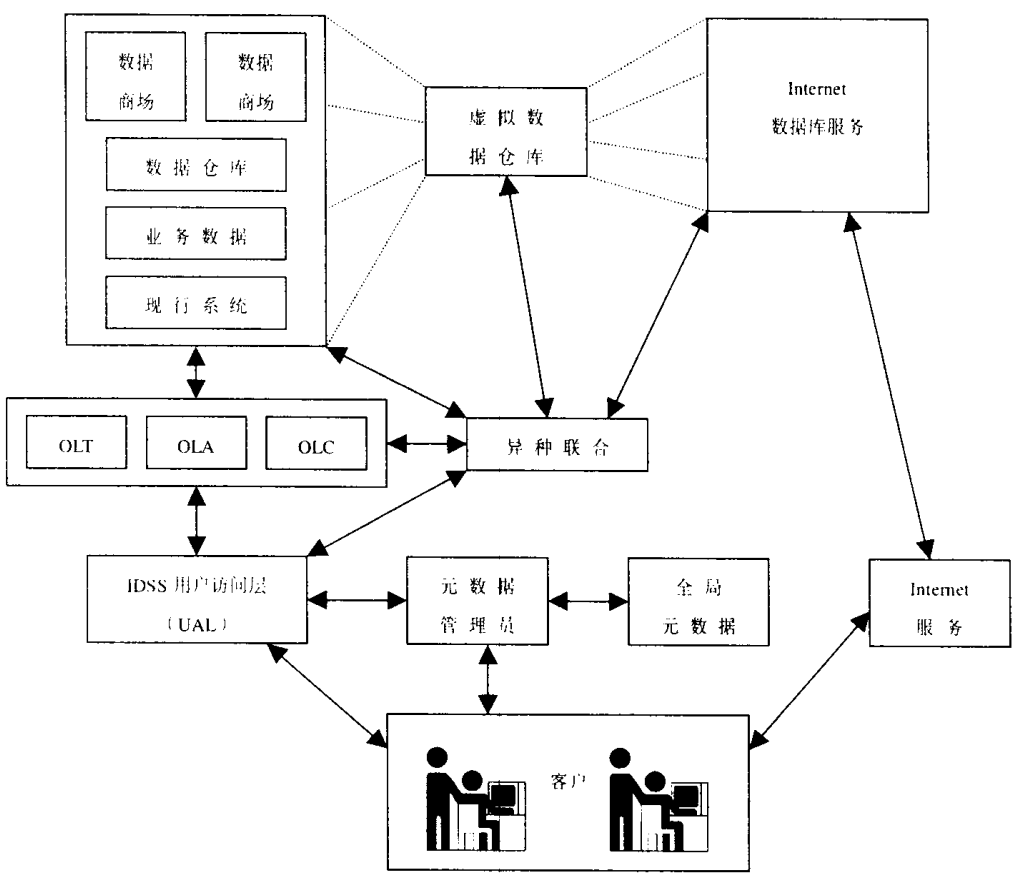


图 1.1　用户角度的 IDSS 或超级数据仓库

数据仓库技术

有必要定义基本术语。约翰·拉得利[6]曾将数据仓库技术定义为由方法、技术和工具所共同构成的在集成的平台上为最终用户提供数据的手段。数据仓库技术不是数据仓库，它只是用来创建存储在传统服务器、 Intranet 服务器或 Internet 服务器上的数据仓库、业务数据商店或数据商场的。技术本身没有什么特别之处，问题在于如何应用技术来解决业务问题。

什么是数据仓库

虽然决策支持系统已经存在了多年，但到九十年代初才将其称为数据仓库，并使高层经理注意到其业务潜力。不同的人对数据仓库有不同的理解。有些定义只局限于数据，有些则指人、处理、软件、工具和数据。虽然确切的措词不同，但都有共同点。W.H. 因蒙[4]将数据仓库明确地定义如下:

> 数据仓库是集成的面向主题的数据库集合，它是用来支持决策支持功能的，其中每个数据单位都与时间相关。

这些数据应该是良好定义的、一致的、不变的。另外数据量也应足够支持数据分析、查询、报表生成和与长期积累的历史数据的对比。数据仓库就是一种能满足上述所有目标的方法。它不是一种能买到的产品，必须循序渐进地构建。

每个行业的许多领域都有成功的数据仓库。一项非正式的调查表明有许多支持销售和营销职能的数据仓库。每家公司都要了解顾客的人口统计特征。除了了解现有顾客外，还想知道什么人没有买他们的产品。这类分析需要的数据经常是从外部来源购买并加入到数据仓库的。这时数据仓库就可用来为营销目的提供潜在顾客的名单。还有提供客户追踪、财务分析、索赔、欺诈分析及其他用途的数据仓库。数据仓库必须提供支持合并业务数据、信息数据、外部数据、部门数据和个人数据框架。就象建楼需要设计图一样，构建能发展成 IDSS 的数据仓库也需要技术框架。

由于许多公司在分析顾客信息时只使用平均数，有些公司就怀疑是否需要数据仓库中的细节数据。虽然初步分析能提供重要的信息，但为获得带来竞争优势所需的信息的话，还需要细化展示细节数据。要注意的是，普通公司只使用平均数，而超级公司则是靠注重细节数据才脱颖而出的。他们需要用不同的方式多层次/多视角查看数据，从而更深入地了解顾客和其他数据。

数据仓库必须是用户驱动的。有些公司曾试图不让用户参与来构建数据仓库，但最终失败了。他们以为从用户已提出的现有要求的基础上就能够了解用户的全部需要，但不幸的是，许多用户是因其要求未及时满足而不再要求新的数据和功能。现有的要求只是冰山一角。

需要单独的环境

为什么数据仓库要求单独的环境呢？为什么决策支持功能无法在业务环境中运行呢？虽然现有的业务能够支持日常的业务职能，但还是需要数据仓库来支持分析、决策和信息用途。现行的业务环境在很多方面都有不足之处。如果能应用速度无限的硬件和效率奇高的软件，那么（理论上来说）只需要一个环境。不幸的是，实际的软、硬件的局限性妨碍了在一个环境中同时支持业务目标和决策支持目标。

另外现行环境中数据的类型、数量和质量通常都无法支持决策工作。数据仓库常用于分析长期趋势。业务系统只包含满足日常业务要求所必需的数据，而数据仓库则包含满足信息要求的大量的历史数据。数据仓库中的数据必须一致、集成、良好定义且有时间标记，而现行环境在这方面就无能为力了。而且特别查询时需要合并业务数据、外部数据和个人数据，大多数业务系统无法实现这一目标。

另外访问的类型也不同。业务环境下用户知道他们需要什么，只是访问单个事务中几行数据。在数据仓库中，为回答一个简单的查询可能要检索成千上万条记录，所以数据的移动量极大；而看到答案后，用户可能会决定再用另一种方法来分析数据。一位备受挫折的数据库管理员抱怨数据仓库的用户“自己都不知道他们想要什么，我怎么能提供呢？”这是个很好的问题。其答案是提供一种灵活的、带有元数据与基础结构的数据体系结构和数据库设计

来支持用户需求的迅速变化。

什么是业务数据商店

业务数据商店（ ODS ）是业务处理的基础，可向数据仓库馈送数据。因蒙、因霍夫和巴塔斯[5]将其定义为一个具有以下特点的体系结构:

- 面向主题
- 集成
- 多变
- 目前很有价值
- 仅包含公司的细节数据

ODS 中的数据是围绕诸如顾客、产品、订货、保险单、索赔等主题的。它是由现行系统提供的，包含整理、转换或集成的数据。多个应用软件都可共享这些数据，而更新数据仅在 ODS 一处完成。对于想把现行系统移入集成环境的组织来说， ODS 是十分有效的。一家大型电讯公司正将成百个现行文件中的顾客信息合并进单个共享的 ODS 中，公司将这个 ODS 称为数据仓库，因为他们是将其用于业务和信息目的。通常的定义认为， ODS 包含的是用于业务目的的数据，而数据仓库包含的才是用于信息目的的数据。由于技术的进步， ODS 和数据仓库间的界限日趋模糊，许多组织开始以上面这种方式来使用数据。 ODS 是某个组织所用的体系结构中的可选结构。许多公司采用的另一种方案是将细节数据直接移进数据仓库，不允许在联机环境中更新数据。

这两种环境的差异如图 1.2 所示。仔细考察图示后可以看出两种环境间存在重大差异。数据仓库的工作单元往往要比业务环境的工作单元大得多，频率也无法预知。因此，在繁忙的生产环境中，很多用户进行特别查询造成的影响是难以接受的。软、硬件的提高已经允许对多个环境进行某种程度的合并，因此最好还是把上述定义当作是指导，而不是僵硬的法规。

项目	业务数据商店	数据仓库
建立方式	现行环境中，每次一个应用软件; 或在 ODS 中，每次一个主题领域。	每次一个或多个主题领域
需求	已知	模糊
关键用途	日常业务操作	影响利润的管理决策
数据访问	单个调用中，检索少量的行	检索需要扫描大量数据
调整方向	对少量数据的频繁访问	对大量数据的非频繁访问
数据量	日常业务所需的数据量	支持统计分析、预报、特别报表和查询所需的大数据量
数据保留	数据保留要满足日常需求	数据保留时间较长，要支持历史报表、对比、分析等
数据时效	即时	通常表示一个静态的时刻；重要的是数据不会随时变化
数据有效性	很高的可用性	除非需要进行世界范围的访问，否则不必具有和生产环境一样的可用性。
工作单元	小、可控制、可预测	大、不可预测、经常变化
设计重点	性能	灵活

图 1.2　业务数据商店与数据仓库的特点比较

数据商场

开发商曾大肆鼓吹过数据商场，它是满足特定用户群需要的数据仓库。随主题领域的不同，可大可小。规划、设计和实施方面的问题都与数据仓库对应阶段中的问题完全相同。元数据、数据体系结构、数据整理、一致性、完整性、可访问性和管理方面的问题也和数据仓库遇到的完全相同。既然要考虑的问题基本相同，本书就不再将数据商场作为独立的问题加以讨论了。设计数据商场时，可以考虑也可以不考虑公司的可访问性和标准。要注意：数据商场中的数据早晚会引起别的用户的兴趣。数据商场建成后，就应该考虑公司的硬件、软件、

网络、数据库管理系统和命名标准。有些开发商和咨询顾问喜欢绕开信息技术人员，向最终用户直销，不让 IT 人员参与实施。这时数据商场就变成了昂贵的信息孤岛，公司中其他人无法访问。即使在没有 IT 人员参与的情况下开发出数据商场，公司利益也要求数据具有广泛的可访问性。

开发商发现直接向最终用户销售数据商场容易成交。公司的 IT 部门对此应该有所察觉，并努力在实行数据商场的所有阶段中直接与开发商和用户合作。数据商场是不可能孤立的，别人早晚都会要求访问本地数据商场中的数据。如果它实施时没有遵从 IT 标准，数据就很难访问。数据商场的建立必须基于公司的硬件、操作系统、网络、数据库、命名标准和元数据等标准，这一点至关重要。

单个组织的数据流

图 1.3 所示为许多公司中都可见到的跨数据商店的数据流。现行系统的数据在移入业务数据商店、数据仓库或数据商场前必须进行整理、转换或集成工作。 ODS 、数据商场或数据仓库的构建顺序是可变的，依赖于业务需求。应用数据仓库技术最好的起点是数据商场。出于对性能的考虑，很多公司在 ODS 中只保留少量的数据。数据仓库可包含也可不含细节数据。对 ODS 、数据仓库和数据商场的经典定义中存在一个问题，即细节数据可冗余存放在上述三种数据商店中。对超大型数据库来说，冗余储存的代价会大得惊人。某些公司（如前文讲到的大型电讯公司）倾向在 ODS 中存储细节数据，在数据仓库中存储概括数据。 ODS 、数据仓库或数据商场可向个人数据仓库提供数据。个人数据仓库中的数据也是个人的，其可用性和有效性是由用户控制的。虽然可自动生成个人数据仓库的备份，但 IT 人员不对其负责。未经 IT 处理确保其完整性，就不能允许它流回其他结构中。

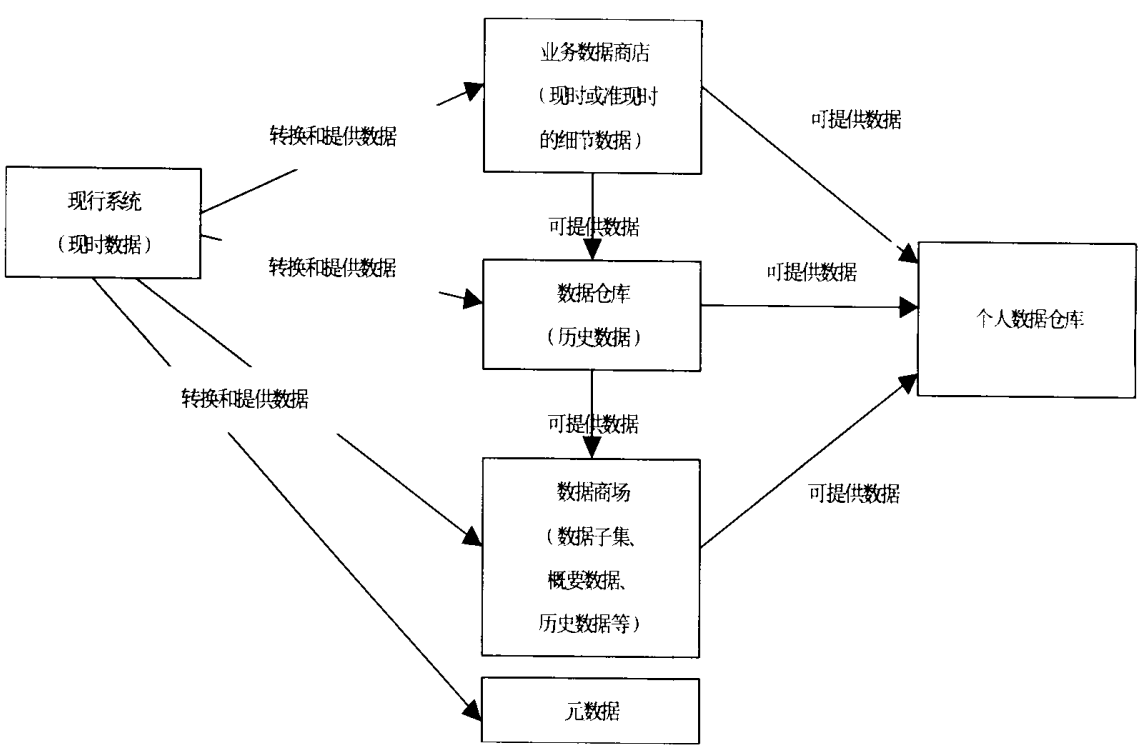


图 1.3　单个组织的数据流

需要元数据

如果能满足用户对信息的需求，他们的能力就会大大增强。这要求他们知道现有什么数据、在何处、如何访问。现行的业务系统同样力不能及。数据分析员多年来一直提倡使用数据字典对数据库进行正确的设计并编写文档。不幸的是，大多数组织并没有接受这一意见，结果现行系统中大量数据都没有文档。无论提供了多么灵活的查询工具，单靠工具本身也解决不了问题，还需要有文档完备的元数据。

小结

本章描述了应当作为数据仓库开发基础的所谓信息数据超市或超级数据仓库的体系结构，并定义了此框架内的各种结构，提出在开发业务数据商店和数据仓库时遵从这一框架在战略上的重要性。数据仓库与其说是终点，还不如说是一段旅程，要想到达那个难以捉摸的终点，了解道路是至关重要的。

参考书目

1、 Bischoff, J. and Yevich, R., "The Superstore: Building More than a Data Warehouse." *Database Programming and Design*, September 1996.

2、 Codd, E.F., "A Relational Model for Large Shared Data Banks." *CACM*, 13, No. 6, June 1970.

3、 Codd, E.F., Codd, S., and Salley, C., "Providing OLAP (On-line Analytical Processing) to User-Analysts: An IT Mandate," E.F. Codd & Associates, 1993.

4、 Inmon, W.H., *Building a Data Warehouse*, QED Technical Publishing Group, 1992.

5、 Inmon, W.H., Imhoff, C. and Battas, G., *Building the Operational Data Store*, John Wiley and Sons, New York, 1996.

6、 Ladley, John, "Operational Data Stores: Building an Effective Strategy," *Data Warehouse: Practical Advice From the Experts*, Prentice Hall, Englewood Cliffs, NJ, 1997.