普通高等教育“十一五”规划教材
PUTONG GAODENG JIAOYU SHIYIWU GUIHUA JIAOCAI

SHUJU CANGKU YU SHUJU WAJUE
YUANLI GONGJU JI YINGYONG

数据仓库与数据挖掘
原理、工具及应用

潘 华 项同德 编著

中国电力出版社
http://jc.cepp.com.cn

普通高等教育“十一五”规划教材
PUTONG GAODENG JIAOYU SHIYIWU GUIHUA JIAOCAI
电力企业信息化系列教材

SHUJU CANGKU YU SHUJU WAJUE YUANLI GONGJU JI YINGYONG

数据仓库与数据挖掘原理、工具及应用

电力企业信息化系列教材编委会

主　任　周光耀
副主任　施泉生
委　员　王乐鹏　潘　华　张科伟　张建华
　　慈向阳　秦天保　张世翔　李　妍
　　赵文会　范忠骏　崔树银

本书编著　潘　华　项同德
主　　审　王翠茹

中国电力出版社
http://jc.cepp.com.cn

内 容 提 要

本书为普通高等教育“十一五”规划教材，是电力企业信息化系列教材之一。

本书全面深入介绍了数据仓库、联机分析处理（OLAP）和数据挖掘的基本概念、工具及实际应用。全书分成三篇，数据仓库与数据挖掘原理篇的主要内容包括数据仓库的基本概念和结构、创建过程、联机分析处理、数据挖掘的基本概念和方法等；数据仓库与数据挖掘工具篇介绍几个现在市场上主流的数据仓库和数据挖掘工具，包括ETL工具Data Stage、商务智能工具Congos和数据挖掘工具SAS；数据仓库与数据挖掘应用篇以某电力公司为例介绍一个数据仓库应用系统的建设过程，包括系统需求分析、系统架构设计、数据模型设计、数据库规划、ETL开发等。

本书可作为计算机、信息管理与信息系统等相关专业的学生学习数据仓库、OLAP及数据挖掘技术的实用教程，也可供从事数据仓库、数据挖掘研究、设计、开发等工作的科研人员和工程人员参考。

图书在版编目（CIP）数据

数据仓库与数据挖掘原理、工具及应用/潘华，项同德编著．北京：中国电力出版社，2007

普通高等教育“十一五”规划教材

ISBN 978-7-5083-6310-3

Ⅰ．数… Ⅱ．①潘…②项… Ⅲ．①数据库系统－高等学校－教材②数据采集－高等学校－教材 Ⅳ．TP311.13 TP274

中国版本图书馆CIP数据核字（2007）第183565号

中国电力出版社出版、发行

（北京三里河路6号 100044 http://jc.cepp.com.cn）

北京丰源印刷厂印刷

各地新华书店经售

*

2007年12月第一版 2007年12月北京第一次印刷

787毫米×1092毫米 16开本 17印张 418千字

定价 **27.00**元

前　言

为贯彻落实教育部《关于进一步加强高等学校本科教学工作的若干意见》和《教育部关于以就业为导向深化高等职业教育改革的若干意见》的精神，加强教材建设，确保教材质量，中国电力教育协会组织制订了普通高等教育“十一五”教材规划。该规划强调适应不同层次、不同类型院校，满足学科发展和人才培养的需求，坚持专业基础课教材与教学急需的专业教材并重、新编与修订相结合。本书为新编教材。

自20世纪90年代以来，信息技术在我国电力系统的应用得到了前所未有的发展，各级电力企业纷纷建立各种各样的信息系统，如办公自动化（OA）、生产管理系统、设备管理系统、燃料管理系统、电力市场和营销系统、电力调度系统、送电和配电地理信息系统、呼叫中心（Call Center）等。然而，这些信息系统往往是根据某个企业，甚至是某个部门自身需求而设计的，信息的采集、加工和存储大多着眼于本企业或本部门的信息，忽视了相互之间信息沟通和共享的要求。这样建立起来的信息系统虽然覆盖了各方面的信息，但同时也形成了一个个信息孤岛，使得原本可以相互沟通和共享的信息被一道道“篱笆”分隔开来。

2002年电力体制改革之后，电力企业解除管制的商业环境以及更加多变的电力市场，使得信息和知识成为电力公司最有价值的资源，而上述情况使得电力企业信息化最终不能构造有效的知识管理系统，信息传递困难，难以提供企业级的决策分析支持。目前的问题主要表现为以下几项。

（1）异构性强，信息集成度差。电力企业各应用系统在数据建模、软硬件平台、应用系统平台和开发工具等方面都存在着显著的差异，从而导致彼此数据交换困难，使得各个应用系统在信息上成为相对孤立的“自动化孤岛”，不易与其他系统交换数据或在企业范围内实现集成。

（2）数据冗余和多信息源问题。由于建设时期的不同以及当时技术水平的限制，造成了过量的数据冗余和多信息源等问题，使得数据资源访问困难，难以进行有效的决策分析。

（3）缺乏企业级的决策支持系统。电力企业各应用系统信息共享困难，管理系统难以跨应用系统实施生产业务流程管理，不能构造有效的知识管理系统，难以提供管理层和决策层的综合分析和辅助决策支持。

数据仓库和数据挖掘技术可以很好地解决以上问题。这种技术自20世纪90年代初开始在美国等国家流行，并在20世纪90年代中期传入我国，现在已经逐渐在我国推广应用，特别是在金融、电信、制造、零售等企业，发挥着越来越重要的作用。相比而言，由于体制、观念、技术、人才等方面的原因，数据仓库与数据挖掘技术在电力行业的应用尚处于起步阶段。但是可以预测，随着电力体制改革和行业信息化的进一步深入，数据仓库和数据挖掘技术将会在电力行业有很大的应用。

本书编者一直从事数据仓库、数据挖掘方面的研究与开发，所参与设计和开发的项目涉及金融、保险、电力等多个领域。近年来在上海电力学院也开设了相关课程。本书是在此基础上编写而成的。与其他此类书籍相比较，本书有自己的一些特色：

（1）基于电力行业应用来介绍数据仓库和数据挖掘的原理。

（2）详细介绍数据仓库主流开发工具架构及其使用。

（3）详细介绍数据仓库在电力行业中的应用现状及相关实例。

全书共三篇，分别是数据仓库与数据挖掘原理篇、工具篇、应用篇。内容组织的思路为：基本概念→基本原理→开发工具→实际应用。

本书在内容介绍上力求深入浅出，通俗易懂。除理论联系实际外，还使用了大量的图示及实例，使得该书有较强的可读性和可理解性。因此，凡具有一定数据库基础知识的人都能学会本书的内容。

本书适合于企业信息化管理人员、技术人员以及软件开发人员阅读，也可作为在校大专、本科学生和研究生的教材。

本书的写作过程也是编者学习、研讨、提高的过程。在此过程中，编者参考了大量网站和图书资料，特别是参阅和引用了不少前辈和同行的工作成果，是他们的一些工作成果使得本书能够比较系统、全面地反映一些有关数据仓库和数据挖掘方面的最新研究成果。书中所引用部分作者的研究成果已经在参考文献中列出，在此表示衷心的感谢！本书得到了上海电力学院的大力支持和帮助，特别是得到了上海市电力经济与管理本科教学高地的资助，在此表示诚挚的感谢！

本书由潘华、项同德编著，其中潘华编写第一、三篇，项同德编写第二篇。本书由潘华担任统稿工作，由华北电力大学王翠茹教授担任主审。

特别感谢上海电力学院党委书记周光耀教授亲自为本丛书作序。

由于我们水平有限，书中难免有疏漏和不妥之处，恳请各方面专家、学者及广大读者批评指正。编者的电子邮箱：panhua@shiep.edu.cn。

编者

2007 年 8 月

电力企业信息化系列教材序

电力行业是技术密集和装备密集型产业，独特的生产与经营方式决定了其对企业信息化的迫切需求。电力企业信息化是指信息技术在电力工业中的应用，是电力工业在信息技术的驱动下由传统工业向高度集约化、高度知识化、高度技术化工业转变的过程。

我国电力行业信息化起步早。20 世纪 60～70 年代，电力行业首先开展了生产、调度自动化的应用，20 世纪 80 年代后期逐渐开展了企业管理信息化的建设。随着科学技术的发展，电力工业正在向高度集约化、高度知识化、高度技术化的工业迈进，电力企业投入巨资引进和开发了如能量管理系统（EMS）、SCADA、自动发电控制（AGC）、电能计量系统、市场预测和分析系统、电力期货交易和短期交易系统、电厂报价决策支持系统、企业资源计划（ERP）、企业资产管理（EAM）系统、电厂监控信息系统（SIS）等。

总体来说，电力企业信息化处于较高水平，但生产自动化与管理信息化的发展处于不平衡状态。一方面，管理信息化滞后于生产自动化；另一方面，生产自动化系统与管理信息系统处于相互分离状态，彼此不能有效互通。这主要是由于电力系统对生产安全性、稳定性和可靠性的要求，导致电力企业对生产过程控制的信息技术应用一向比较重视，而对业务及管理的信息化重视却相对不足，遗留下很多待解决的问题，如企业信息资源整合，数据共享性、同一性问题，企业统一信息平台，信息系统的标准化，信息安全以及企业信息化的体系架构，信息编码标准化和交换规范等。

作为国内首部公开出版的电力企业信息化系列丛书，本丛书试图以电力工业发、输、配、供四大环节核心价值链为主线，构建电力企业信息化整体框架模型，并在该整体模型的基础上建立围绕不同类型的企业、不同的信息化应用层次建立若干个专题，以求全面把握电力企业信息化体系和功能，为解决电力企业信息化遗留问题提供一些思路。

上海市教委在深入研究分析上海高等教育各层次、各学科的教育资源现状的基础上，依据国家对上海近、中期人才培养的要求，依据上海建设“四个中心”的国家战略以及上海对先进制造业、现代服务业的需求，于 2005 年启动了“上海高等学校本科教育高地建设”项目，确定了金融保险、海关物流、外贸经济等十大高地，采用“项目申报制”，每年投入专项资金重点建设。目标是将上海高等学校建设成为上海乃至全国的人才培养重要基地和高等学校教学研究与师资培训中心，成为在国内外有一定知名度和影响力的本科教育高地，为上海城市发展和经济建设提供人力资源保障。目前，我院承担上海市电力经济与管理本科教育高地建设，明确了要与上海优先发展先进制造业和现代服务业的战略要求相适应。本丛书反映了本院本科教育高地建设的要求，是本科教育高地建设的成果之一。

本丛书计划编撰和出版六本，分别是《电力企业信息化概论》、《发电企业信息化及案例分析》、《电网企业信息化及案例分析》、《电力企业决策支持系统原理及应用》、《电子商务原理及应用》、《数据仓库与数据挖掘原理、工具及应用》等。本丛书的作者长期从事电力经济

管理的教学工作，积累了大量的典型应用案例，注重理论分析和典型应用案例相结合，既具有理论深度又具有可理解性和可操作性，也是本书的鲜明特点。本丛书中，电力工业各环节信息化与整体框架紧密承接，又自成体系，既能够满足本科教学，又能作为行业培训教材使用，满足不同层次、不同需求的读者需要。

上海电力学院党委书记 周光耀

2007 年 5 月

目　录

第二篇　数据仓库与数据挖掘工具篇

第三篇　数据仓库与数据挖掘应用篇

第一篇　数据仓库与数据挖掘原理篇

第一章　数据仓库概述

第一节　数据仓库的产生

信息技术在企业生产各个环节的深入运用极大地改善了企业的业务流程，提高了企业的生产效率。企业的生产经营活动由从事企业生产过程的业务活动和进行企业经营决策的分析活动组成。与之对应的信息系统同样分成了两种类型：一种是面向业务操作层面的应用系统，专注于规范企业的业务流程；另一种是面向决策分析层面的应用系统，专注于提高企业的决策水平，指引企业的发展方向。我们分别把这两种系统称为联机事务处理（On-Line Transaction Process，OLTP）系统和联机分析处理（On-Line Analytical Process，OLAP）系统，因此数据处理也相应地划分为两大类：操作型数据处理和分析型数据处理（或信息型处理）。

联机事务处理系统是随着数据库技术的不断完善发展起来的，典型的系统如企业资源计划（ERP）系统、供应链（SCM）系统、面向很多行业的交易系统以及各种类型的管理信息系统（MIS）。联机事务处理系统体系架构也由早期的单机系统发展到客户—服务器（Client/Server，C/S）架构的两层架构，随着网络技术的不断完善，随后又出现了基于 Web 的三层和多层架构。联机事务处理系统面向业务活动渗透到企业运作流程的各个环节，及时详细地记录下企业运行中的各种数据。随着时间的推移，联机事务处理系统中积累了大量的详细数据，并且被分散存储在不同的计算机系统中。由于企业的业务系统可能由不同的厂商在不同的时间分阶段开发完成，并且为了实现不同的业务目标和出于系统安全等多方面因素的考虑，这些系统相互之间并不沟通。我们称这些孤立在某些计算机系统中的数据为"信息孤岛"。信息孤岛中的数据量越来越大，种类越来越多，结构也越来越复杂。使得企业不能充分利用和管理信息资源，出现了"企业数据泛滥，信息匮乏"的尴尬局面。对一些来自高层或管理人员的数据分析需求常常表现得无能为力。下面是某电力公司的一个例子。

W 先生是某省电力公司总经理，该公司在信息化建设中投入了大量的资源，建立了财务、呼叫中心、电力负荷管理、人力资源管理等业务系统。这些业务系统都非常先进、高效。可以说，企业的很多数据都可以在这些业务系统中找到，W 先生也为此感到非常自豪。然而，从几年前的某一天开始，W 先生几乎麻烦不断，以下是他的部分经历。

（1）一天，他坐在办公室，想了解不同地区、不同行业、不同电压等级、不同时间、几个大客户的用电情况，却被告知，要获得这样的信息很困难，这使他无法理解，"为什么这些数据都在计算机里，而我却看不到?"

（2）通过不同的途径计算出来的客户信用度总是不一致，不是多一点就是少一点，很难确认哪一个是正确的。

（3）他要出席一个重要会议，突然打电话要求几张欠费分析情况的临时报表，其中包括欠费对企业流动资金、利润的影响等几项复杂指标。由于这些信息存在于不同的系统中，结果，W先生和他的部下经过几天的奋战，终于在最后1min把报告交到了上级主管手里。现在他只能祈祷千万别因为时间紧迫而在报告中有什么差错，并且以后这样的会议尽量少一点。

（4）一些系统如统计分析、客户发展潜力分析等总是需要用到其他系统所产生的数据，而这些要求往往涉及不同部门间的协调和不同系统间的数据交换，这让W先生非常头疼。

（5）公司计划建立一个庞大的客户服务分析系统，而这个系统几乎需要使用到全企业的所有数据，还要用到之前10年的很多数据，在当前的数据环境下，这几乎是不可能的。

这些问题出现的根源在于分析型数据处理系统的缺失。该电力公司目前建立的所有系统都是事务处理系统，是面向操作型数据处理的，系统设计的初衷是满足工作人员的日常操作，即对一个或一组记录的查询和修改，主要为企业特定的业务流程服务，用户关心的是响应时间、数据的安全性和完整性。例子中列举的所有工作已经超出了事务处理系统支持的范围。要想从根本上解决这些问题，企业必须建立面向经营决策的分析型数据处理系统。数据仓库（Data Warehouse，DW）也就是在这种背景下产生的。数据仓库是一种新的数据处理体系结构，它是对企业内部各部门业务数据进行统一和综合的中央数据仓库。它为企业决策支持系统和行政信息系统提供所需的信息。它也是一种信息管理技术，为预测利润、风险分析、市场分析以及加强客户服务与营销活动等管理决策提供支持。

数据仓库作为一种从数据库发展而来的新型技术，用于分析型处理，它弥补了许多传统数据库的不足之处，其最大的用途是提供给决策者一种全新的方式，从宏观或微观的角度来观察多年累积的数据，从而使决策者可以迅速地掌握自己企业的经营运作状况、运营成本、利润分布、市场占有率、发展趋势等对企业发展和决策有重要意义的信息，以利于做出更加及时、准确、科学的决策。

操作型处理和分析型处理的分离划清了数据处理的分析型环境与操作型环境之间的界限，使企业数据环境由原来的以单一数据库为中心的数据环境发展为以数据库和数据仓库为中心的企业数据环境，如图1-1所示。

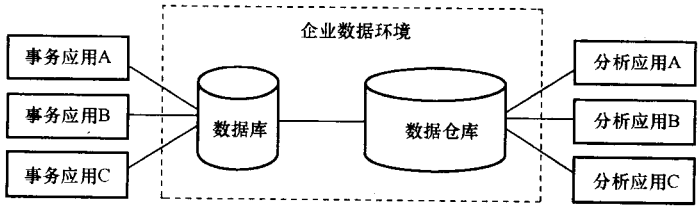

图1-1 以数据库和数据仓库为中心的企业数据环境

第二节 数据仓库的相关概念

数据仓库的概念是由W. H. Inmon（被称为数据仓库之父）在《建立数据仓库》（*Building the Data Warehouse*）一书中提出的。随着人们研究和行业发展的不断深入，相继涌现出了很多相关的概念，包括数据集市、商务智能等。下面依次介绍数据仓库领域常见的概念。

一、数据、信息和知识

数据仓库的所有领域都是围绕着数据、信息和知识展开的，它们是以数据仓库为基础的企业信息工厂的原料和产品。

传统的数据可定义为客观事物记录下来的、可以鉴别的符号（文字、字符串等），是客观事实的属性、数量、位置及相互关系等的抽象表示。随着科学技术的发展，人们可以使用电子化文档、图像、声音、视频进行交流和记录客观事实，我们把这些新的记录客观事实的形态也称为数据。数据依据组织方式可以分成结构化数据和非结构化数据。结构化数据是按照特定的规则组织起来的数据，这种数据可以通过技术手段方便地访问，典型的例子是存储于关系型数据库中的关系型数据。非结构化数据是松散的、无规则的数据，数据的内容也是没有结构的，这种数据通常的技术无法直接利用。数据通常包含两方面的属性：一是数据的类型，如数值型、字符型、日期型、二进制类型等；二是数据长度，如数值型数据的精度、字符型数据的字符数、二进制数据的字节数等。

信息是人们对数据进行系统地采集、组织、整理和分析的结果，是经过加工以后并对客观世界产生影响的数据。人们一般都是将数据转化成信息的形式加以使用。有的专家认为信息是数据和上下文的结合，可用于决策。一般信息都可以用一组词及其值来描述。

关于知识，人们有很多不同的理解和定义。比较有代表性的知识定义如：Feign 认为知识是经过削弱、塑造、解释、选择和转换了的信息；Bernstein 定义知识是由特定领域的描述、关系和过程组成的；Hayes-Roth 则认为知识＝事实＋信念＋启发式。还有些专家认为知识是经验、价值、有条理的信息、专家的见解和本能的直觉的结合体，它可以提供评价和吸收新的经验和信息的环境和架构。虽然关于知识没有统一的定义，但一般将事实、规则、模式、规律和约束等看作知识。事实是指人类对客观事物属性的值或状态的描述，可以用一个值为真的命题陈述或一种状态的描述来表达。规则可以分为前提条件和结论两部分，用于表示因果关系的知识。如果规则中含有可以实例化为不同具体值的变量，则这种规则称为规律。模式是指符合事物生存运行的内在规律，具有正确的发展导向和行为要求的统一式样、运行机制、管理体制、解决方案的综合，一般可以作为范本、摹本和变本的式样。广义上，知识是类别特征的概括性描述。根据数据的微观特性发现其表征的、带有普遍性的、较高层次概念的、中观和宏观的知识，反映同类事物的共同性质，是对数据的概括、精炼和抽象。知识是以多种方式把一个或多个信息关联在一起的信息结构。

二、数据仓库的概念

数据仓库一词尚没有一个统一的定义，最早是由 W. H. Inmon 提出来的：数据仓库是一个面向主题的、集成的、相对稳定性的、反映历史变化的数据集合，用于支持管理决策。此后，不同的学者从不同的角度为数据仓库下了不同的定义。

Informix 公司负责研究与开发的公司副总裁 Tim Shelter 把数据仓库定义为："数据仓库将分布在企业网络中不同信息岛上的业务数据集成到一起，存储在一个单一的集成关系型数据库中，利用这种集成信息，可方便用户对信息的访问，更可使决策人员对一段时间内的历史数据进行分析，研究事务发展的走势。"

另外，在由 A. Silberscharz 等发表的《数据库研究：面向 21 世纪的机遇与成就》中把数据仓库定义为："来自一个或多个数据库的数据的备份。"

现在，业内普遍公认的是 W. H. Inmon 的定义。该定义指出了数据仓库面向主题、集

成、相对稳定性、随时间变化这 4 个重要的特征。

（一）面向主题

传统的操作型系统中数据是围绕企业的应用进行组织的，与此相对应，数据仓库中的数据是面向主题进行组织的。所谓主题，是一个归类的标准，每一个主题基本对应一个宏观的分析领域。对于一个电力公司来说，应用问题可能是计量管理、电力负荷管理、财务管理、客户服务等，公司的主要主题可能是销售利润、售电量分析、欠费分析等。有时可以对以上主题进一步分解得到二级主题，如售电量分析可以分为地区用电分析、重点客户用电分析、各类电价分析、用电变化趋势预测等。

在图 1-2 中，显示了一个电力企业的情况。该企业基于传统数据库已经建立了营销数据库、呼叫中心数据库、财务数据库等。其中，营销数据库记录了客户用电量信息，呼叫中心数据库记录了客户咨询、报修、投诉等信息，财务数据库记录了客户电费缴纳情况信息。如果要在原有数据库的基础上分析客户电费缴纳、用电量、咨询等全方位的信息无疑是费时费力的，因为这个工作要同时访问 3 个数据库。同样，如果要分析公司收益情况，也会遇到同样的问题。

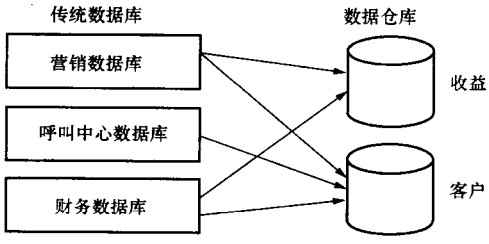

图 1-2　数据仓库面向主题特性

基于以上原因，我们以收益和客户为主题建立数据仓库。其中，收益主题可以从财务和营销数据库中了解公司收入情况，客户主题可以从财务、营销数据库和呼叫中心数据库中获得客户用电量、电费缴纳、咨询等全方位的信息。由此可见，数据仓库极大地方便了数据分析的过程。

（二）集成

在数据仓库的所有特性之中，这是最重要的。数据仓库中的数据是从多个不同的数据源传送来的，当这些数据进入数据仓库时，需要进行转换、重新格式化、重新排列以及汇总等操作。这是因为操作型环境下的数据并不适合用来做分析，做分析时也不需要用到全部业务数据。另外，数据仓库中每一个主题所涉及的数据分散在不同的数据库中，且有很多重复和不一致。图 1-3 说明了当数据由面向应用的操作型环境向数据仓库传送时所进行的集成。

当数据进入数据仓库时，要采用某种方法来消除应用层的许多不一致性。例如，在图 1-3 中，考虑关于“性别”的编码，在数据仓库中数据是被编码为 m/f 还是 1/0 并不重要，重要的是，无论方法或源应用是什么，在数据仓库中应该一致地进行编码。如果应用数据编码为 X/Y，当其进入数据仓库时就要进行转换。对所有的应用设计问题都要考虑同样的一致性处理，比如命名习惯、关键字结构、属性度量单位以及数据物理特点等。

（三）相对稳定性

操作型环境下一般只存储短期的数据，并且数据会随着业务的进行不断地被更新。另外，数据的访问和处理一般按一次一条记录的方式进行。数据仓库中的数据呈现出一组非常不同的特性，其数据通常是以批量方式载入与访问的，而且数据仓库环境中并不进行一般意义上的数据更新，数据仓库中的数据在进行装载时是以静态快照的格式进行的。当产生后继变化时，一个新的快照记录就会被写入数据仓库。这样在数据仓库中就保存了数据的历史状况，所以对于访问数据仓库的最终用户而言，数据是只读的，如图 1-4 所示。

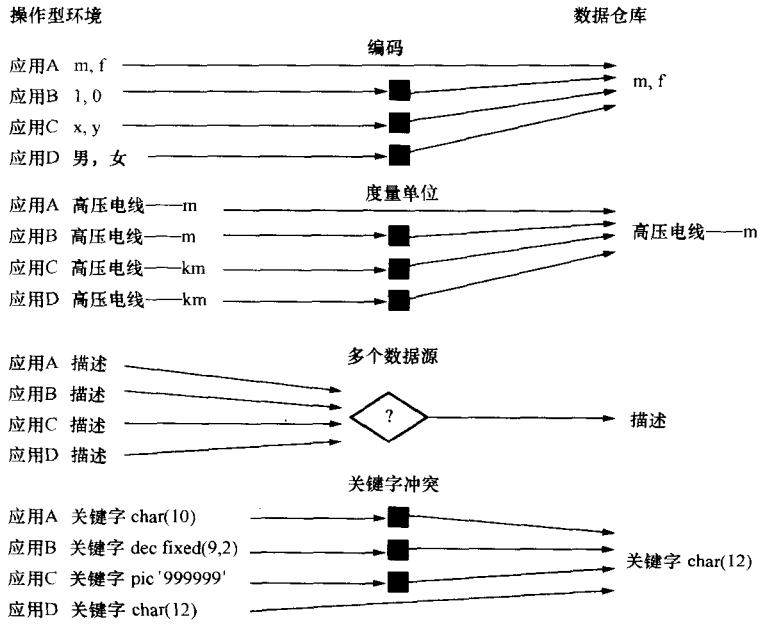

图 1-3 数据由面向应用的操作型环境向数据仓库传送时所进行的集成

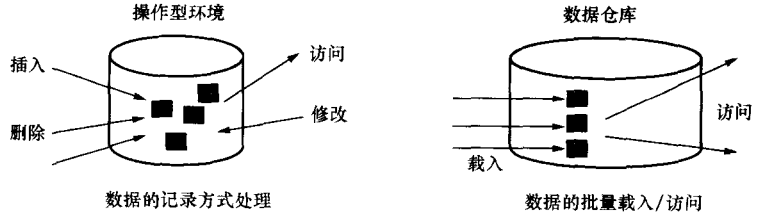

图 1-4 数据仓库相对稳定性示意

（四）随时间而变化

数据仓库中的数据一般按照一个固定的时间间隔批量载入，是稳定的，这使得数据仓库中的数据总有一个时间维度，用来表明数据的历史时期。

操作型环境和分析型环境下有不同的时间范围。数据仓库中的数据时间范围要远远长于操作型系统中的数据范围。操作型系统的时间范围一般是 60～90 天，而数据仓库中数据的时间范围通常是 5～10 年。由于这种在时间范围上的差异，数据仓库含有比任何其他环境中都多的历史数据。

事实上，数据仓库除了以上 4 个重要特性外，还有数据量大、对软硬件要求高等特点。

三、数据集市

数据仓库的工作范围和成本常常是巨大的。针对所有的用户并以整个企业的眼光对待任何一次决策分析，将形成代价很高、时间较长的大项目。因此更紧凑集成的、拥有完整图形接口且价格更具吸引力的工具即数据集市（Data Mart）应运而生。目前，全世界对数据仓库总投资的一半以上均集中在数据集市上。

数据集市是一种更小、更集中的数据仓库，是为企业提供分析商业数据的一条廉价途

径。它是具有特定应用的数据仓库，主要针对某个具有战略意义的应用或具体部门级的应用，它支持客户利用已有的数据获得重要的竞争优势或找到进入新市场的解决方案。

因此数据集市可以看作整个企业数据的一个子集，包括特定业务单元、部门或用户集的值。该子集包含从事务处理或企业仓库获取的历史数据、汇总的数据，并可能会有一些详细数据。数据集市是根据特定主题而不是根据数据集市数据库的大小来定义的。数据集市通常服务于单个部门或企业的部分用户，满足部门级用户的需求，因此数据集市也被称为部门级数据仓库。

数据集市可以分成两种，一种是从属数据集市，另一种是独立数据集市。从属数据集市的数据直接来自于中央数据仓库，一般是为那些访问数据仓库十分频繁的关键业务部门建立的，从而很好地提高查询的反应速度。由于基于统一的中央数据仓库，从属数据集市可以很好地保证数据的一致性。独立数据集市的数据直接来源于业务系统。许多企业在规划数据仓库时，往往出于投资等方面的考虑，最后建成的就是这种结构的独立数据集市，用来解决个别部门比较迫切的决策问题。它和企业级数据仓库除了在数据量和服务对象上有所区别外，逻辑结构是相似的。独立数据集市虽然满足了部门级的决策需求，但由于缺乏企业级的集成，独立数据集市之间容易造成数据不一致。目前，在电力行业尤为如此，由于数据仓库投资巨大，对该行业还是一个比较陌生的事物，对其潜在的价值也认识不足，多数电力公司在打算实施数据仓库时，多选择有一定需求的市场营销部门建立数据集市，然后再进行推广，此时建立的就是一种独立数据集市。独立数据集市是企业在特定发展阶段中为适应特殊需要建立的，往往会随着企业级数据仓库的建立而弃用，被从属数据集市取而代之。不过值得一提的是，独立数据集市在建立的过程中已经与部门应用紧密结合，会对企业级数据仓库的建立提供有力的支持。

独立型数据集市和从属型数据集市的逻辑结构如图 1-5 所示。

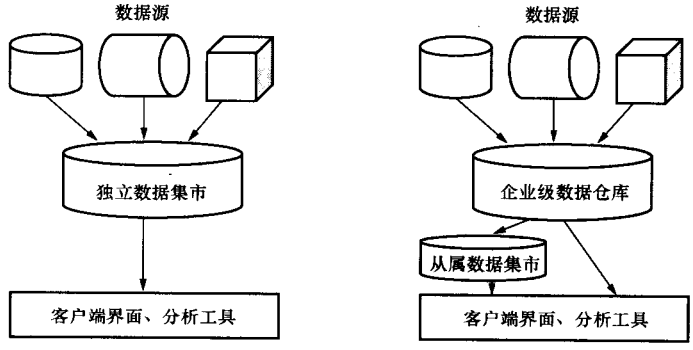

图 1-5　独立型数据集市和从属型数据集市的逻辑结构

需要注意的是，数据集市并不是数据仓库，也不是小型的数据仓库。数据集市是根据用户的功能范围（即特定主题）而不是根据数据集市数据库的大小来定义的。数据集市的累加也不是数据仓库。因为数据仓库的数据覆盖了整个企业范围，并在企业级对数据进行了集成，而数据集市则是分别在部门级对数据进行集成，并没有进行企业级的数据集成，无法提供统一的全局视图。当进行部门级的数据分析时，人们通常通过数据集市提高访问效率，但当分析的问题是企业级的时候，则需要完整的数据仓库。

四、商务智能

在以数据库为中心的业务处理系统和以数据仓库为基础的分析系统的基础上，IBM公司首次提出了商务智能（Business Intelligence，BI）系统的概念，商务智能系统将信息转化为知识，并强调在正确的时间将准确的信息交给合适的用户，从而支持决策过程。TDWI的Wayne Eckerson在*Understanding Business Intelligence*一文中将商务智能系统类比成一座炼油厂——数据炼油厂（Data Refinery），如图1-6所示。

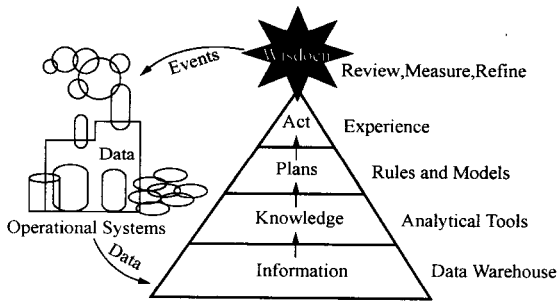

图1-6 将BI环境看作一个“数据炼油厂”

在数据炼油厂中，以数据为原材料，经过数据处理过程，生产出各种能够满足用户特定需求的信息产品，包括信息、知识、计划、行动等。数据处理过程由以下5个部分组成。

（1）从数据到信息的过程：这个过程将业务系统的数据抽取并集成到数据仓库，作为信息存储在数据仓库中。

（2）从信息到知识的过程：业务人员使用分析工具（如查询、报表、OLAP、数据挖掘等）访问和分析数据仓库中的信息，识别信息中存在的趋势、模式和异常，从而使信息转化成业务人员的知识。

（3）从知识到规则的过程：业务人员一旦获取了自己需要的知识后，会从识别的趋势、模式和异常中创建规则和模型，从而指导和改善业务运行。

（4）从规则到计划和行动的过程：在制定了规则之后，人们通过制订计划来执行规则，这样计划便将知识和规则转化成行动。

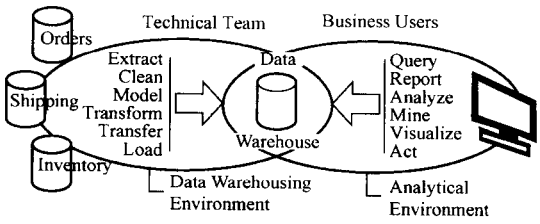

图1-7 包含数据仓库和分析环境的BI环境

（5）循环反馈过程：计划一旦被执行，将反馈到业务系统，知道业务系统的运行，改进企业的流程。

Wayne Eckerson认为一个完整的商务智能系统框架应当包含两个基本环境，如图1-7所示。

（1）数据仓库环境：在该环境中技术人员将花费大量的时间从事抽取、清洗、建模、转化、加载等工作，将一个或者多个业务系统的数据集成到数据仓库环境。技术人员需要熟悉业务环境中大量的数据，这个过程被称为数据考古（Data Archaeology），总结出企业级的单一业务模型，按照业务模型创建数据仓库的框架，并将这些数据集成到数据仓库中。之后可以创建数据集市来满足部门级数据访问需求，提高访问性能。

（2）分析环境：在分析型环境中，业务人员使用分析工具基于数据仓库的数据从事查询、报表、数据挖掘等各种分析活动。技术人员一般根据业务人员的常用需求创建报表，使业务人员能够通过一定的条件访问到自己需要的数据，并且通过争取可以访问到更加细节的数据。为了能够直观地反映计划进度或企业的绩效，可以通过仪表板或平衡记分卡的方式呈现给业务人员。

综上所述可以知道商务智能更加强调企业从数据集成到分析决策的全过程。商务智能系统具有以下主要优点：商务智能系统不仅采用了最新的信息技术，同时提供预先打包好的应用领域的解决方案；商务智能系统着眼于终端用户对业务数据的访问和业务数据的传送，服务于信息提供者和信息消费者。因此商务智能也成了以数据仓库为基础的分析型应用技术的行业代名词。在这个行业当中，数据仓库处在核心位置，同时还要提供方便业务用户查询分析数据的整体解决方案。这些解决方案通常被称为 BI 工具，它们是数据仓库前端开发的核心工具。

第三节 数据仓库与 OLTP 的比较

通过前面的分析，我们可以总结一下操作型数据（OLTP 数据库）与分析型数据（数据仓库）之间的区别，如表 1-1 所示。

表 1-1 操作型数据和分析型数据的区别

操作型数据	分析型数据
细节的	综合的或提炼的
在存取瞬间是准确的	代表过去的数据
可更新	不更新
操作需求事先可知道	操作需求事先不知道
需求驱动的“瀑布式”系统开发方法	由数据开始的“螺旋式”系统开发方法
对响应性能要求高	对响应性能要求宽松
一个时刻操作一个单元	一个时刻操作一个集合
事务驱动	分析驱动
面向应用	面向分析
一次操作数据量小	一次操作数据量大
支持日常操作	支持管理需求

上述操作型数据与分析型数据之间的差别从根本上体现了事务处理与分析处理的差异。传统的 OLTP 数据库系统由于主要用于企业的日常事务处理工作，存放在数据库中的数据也就大体符合操作型数据的特点。但这些数据却并不适用于分析处理，很难直接对它们进行分析。

事务处理环境不适宜分析处理应用的原因主要有以下 5 点。

1. 事务处理和分析处理的性能特性不同

在事务处理环境中，用户的行为特点是数据的存取操作频率高而每次操作处理的时间短，涉及的数据量一般不大，而且数据的更新和查询同样频繁；在分析处理环境中，用户的行为模式与此完全不同，某个分析处理应用程序可能需要连续运行几个小时，涉及大量数据的连接操作，从而占用大量的系统资源。因此，将具有不同处理性能的这两种应用放在同一个环境中运行显然是不恰当的。

2. 数据集成问题

分析处理需要集成的数据，而各个 OLTP 数据库中常常只有与分析主题相关的部分数据。全面而正确的数据是有效的分析和决策的首要前提，相关数据收集得越完整，得到的结